Inferring paths of neoplastic transformation from analysis of colorectal cancer with residual polyp of origin

A THESIS
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Minsoo Kim

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

ALEXEJ ABYZOV, ADVISER
CHAD MYERS, CO-ADVISOR

JUNE 2017

ACKNOWLEDGEMENTS

I extend my sincere gratitude to Dr. Alexej Abyzov for his guidance, patience, and encouragement. I am extremely blessed to have him as my mentor. He always guided me towards stimulating discussions and new ideas. I would like to thank my co-advisor, Dr. Chad Myers. I am grateful for his teaching and continuous encouragement, which have been critical in my academic endeavors.

I want to thank the members of the Abyzov Lab: Taejeong Bae, Dhananjay Dhorkarh, Tanmoy Roychowdhury, Nikolaos Vasmatzis, and Chen Wang. I am grateful for their friendship, teaching, and encouragement.

I would like to thank Dr. Lisa Boardman for the opportunity to work on such an exciting and collaborative research project. I would also like to thank Dr. Nicholas Chia and Brooke Druliner for their collaboration and valuable comments on this thesis. Also, I want to thank my thesis committee member, Dr. Krishna Kalari, for her immense support and insightful comments.

Finally, I would like to thank my parents, Bohyun and Ji-Yun Kim, and my sister, Julia, for their encouragement and love.

ABSTRACT

Besides the classical evolutionary model of colorectal cancer (CRC) defined by the stepwise accumulation of mutations in which normal epithelium transforms through an intermediary polyp stage to cancer, few studies have proposed alternative modes of evolution (MOE): early eruptive subclonal expansion, branching of the subclones in parallel evolution, and neutral evolution. However, frequencies of MOEs and their connection to mutational characteristics of cancer remain elusive. In this study, we analyzed patterns of somatic single nucleotide variations and DNA copy number aberrations (CNAs) in CRC with residual polyp of origin and in cancer free polyps from 27 patients in order to determine this relationship. For each MOE we defined an expected pattern with characteristic features of allele frequency distributions for SNVs in cancers and their matching adenomas. From these distinct patterns, we then assigned an MOE to each CRC case and found that stepwise progression was the most common (70% of cases). We found that CRC with the same MOE may exhibit different mutational spectra, suggesting that different mutational mechanisms can result in the same MOE. Inversely, cancers with different MOEs can have the same mutational spectrum, suggesting that the same mutational mechanism can lead to different MOEs. The types of somatic substitutions, distribution of CNAs across genome, and mutated pathways did not correlate with MOEs. As this could be due to small sample size, these relations warrant further investigation. Our study paves the way to connect MOE with clinical and mutational characteristics not only in CRC but also to neoplastic transformation in other cancers.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Discussion of CRCs

Colorectal cancer (CRC) is the third most common cancer type with respective incidence and mortality rates of 40.7 and 14.8 per 100,000 people [1]. CRC frequently metastasizes to the liver, intestinal lymph nodes, lung, and abdominal cavity [2]. Even though the recent extensive screening and thorough removal of any sighted polyps substantially decreased the incidence rate, studying the causes of the cancer development seems crucial in its prevention and development of appropriate treatments [3].

The foundation for the studies of genetic evolution in colorectal cancer (CRC) was built upon the finding first presented in the seminal work by Fearon and Vogelstein that the accumulation of genetic alterations led to neoplastic transformation in the colon to CRC [4]. This widely accepted and dominant paradigm that CRC arises in a linear model of accumulated genetic mutation and large-scale genomic disruptions of chromosomal instability continues to be the infrastructure upon which extensive research on carcinogenesis is based [5]. In this classical model of CRC, carcinogenesis is presumed to follow a linear trajectory from normal colon tissue to a precancerous lesion, known as an adenomatous polyp, to cancer. Although the large structural variations and copy number aberrations (CNA) along with the typical accumulation of point mutations suggested that the carcinogenesis of colorectal cancer required chromosomal instability (CIN), this speculation was debunked with the discovery of microsatellite Instability (MSI) type colorectal cancer [6]. According to Boland et al., microsatellite sequences are regions of simple repeats that are particularly prone to genetic mistakes due to this repeating characteristic. Naturally, MSI refers to the phenotype of a large number of

mutations within these microsatellite sequences. In his study, Boland states that this is either due to abnormal DNA repair systems, specifically, deficient DNA mismatch repair (MMR) or due to hyper-methylation in MLH1 gene MSI. The colorectal cancer development through MSI is different from the traditionally studied CIN types as the large number of mutations in microsatellite sequences is often accompanied by an absence of CNA. This lack of large structural variation is one of the potential reasons why MSI cases have a better prognosis than CIN cases [7].

In addition to the classification of CRC via developmental mechanisms, hereditary status also can classify colorectal cancer into two types: familial and sporadic. Approximately 30% of the CRC is hereditary, and among these, hereditary non-polyposis colorectal cancer (HNPCC) and familial adenomatous polyposis (FAP) are the two types that are extensively studied [8]. HNPCC, also known as the Lynch syndrome, is a rare genetic disorder with germline variants in DNA MMR [6]. Although it accounts for approximately 4% of all CRC cases, HNPCC can increase the risk of developing adenomas up to 80% [9,10]. Because of their defective DNA MMR, Lynch syndrome is often associated with a high level of microsatellite instability (MSI-H). Sporadic colorectal cancer, on the other hand, refers to non-hereditary colorectal cancer. As Boland explains in his study, these develop due to acquired somatic mutations during the lifetime of the individual and can also develop with microsatellite instability. In fact, it was found that approximately 12% of the colorectal cancers diagnosed have developed MSI through the accumulation of somatic mutations [6].

One of the most extensively studied pathways relevant to colorectal cancer would be the Wnt-signaling pathway. In a normal cell, Wnt extracellular protein binds to its

receptor and initiates a cascade leading to an accumulation of β-catenin proteins [11]. According to the authors, the Wnt-signaling pathway works as follows: β-catenins subsequently are transported into the nucleus, and the recruitment of these β-catenins promotes proliferation. When the Wnt proteins are not present outside the cell, their receptors are not active and cytoplasmic proteins would recruit any excess β-catenins and degrade them within the cell to prevent the translocation of the β-catenins and unnecessary proliferation that is followed. One of the cytoplasmic proteins responsible for maintaining β-catenin at low levels during a normal state is the *Adenomatus Polyposis Coli (APC)* gene [12]. In his study, Markowitz states that *APC* gene is a well-known tumor suppressor because mutations in this gene allow β-catenins to both increase in number and localize to the nucleus, resulting in an uncontrollable proliferation. Inactivation in the *APC* gene has been known to be one of the hallmarks of colorectal cancer progression, especially through a hereditary knock out of one of the alleles that predisposes individuals to a higher risk of developing CRC [13].

Another major mechanism that is well studied in colorectal cancer is the *p53*, a well-known tumor suppressor gene. The role of *p53* inactivation in cancer progression has been observed across multiple cancer types [14,15]. While the normally functioning *p53* acts as a checkpoint to various DNA damage by regulating cell cycle arrest, DNA repair, or apoptosis, inactivation of the gene promotes cells to grow insensitive to these regulating measures. In colorectal cancers, *p53* was reported to be commonly inactivated through the entire short arm deletion of chromosome 17 [16].

Sequencing Technology

While cancer has been characterized clinically and histologically, with a focus on the major genes and specific pathways that contribute to the development of cancer, the revolution in cancer studies came with the advent of sequencing technology. Methods of sequencing DNA strands have existed since the 1970s with the goal of obtaining a complete sequence of small nucleotide fragments [17]. Sanger sequencing was really the first method to produce and sequence small fragments of DNA strands. As with any sequencing technology, Sanger sequencing can be divided into two different types: de novo sequencing and re-sequencing with a reference sequence [18]. As explained by Shendure, de novo sequencing is based on the shotgun method of cutting DNA into random segmentations and transformation of the segments by bacteria for cloning. Once the bacterial colony is grown, the plasmids are isolated and sequenced with ddNTPs that terminate the addition of new nucleotides in a growing strand. The resulting strands were run on the electrophoresis for sequence determination. Re-sequencing was simply performing the polymerase chain reaction (PCR) since the restoration of the original sequence from the segments can be done if it is already known where the segments are supposed to be ligated. After cycles of denaturation of DNA in high temperatures followed by annealing of it in low temperatures and elongation, PCR generates multiple DNA strands accurately. These DNA strands of varying length can be then separated by weight via electrophoresis and the different lanes would indicate the appropriate nucleotide letters in a sequential manner.

Despite the ability of Sanger sequencing to obtain the sequence of DNA strands, nation-wide efforts in obtaining a complete sequence and mapping of our entire genome

through initiatives such as the Human Genome Project was limited by the scalability of production and sequencing of DNA strands needed for the projects. This need for a mass production in DNA sequencing led to the introduction of second-generation sequencing, more commonly known as next-generation sequencing. As opposed to the traditional sequencing of DNA strands floating in aqueous solutions, second-generation sequencing was founded in an array-based technology in which the DNA fragments were cloned in parallel for faster results [19]. With this new array-based technology, hundreds of millions of sequencing reads could be processed simultaneously. This significantly increased the efficiency of sequencing but required an image capture that was prone to error in calling the nucleotide bases, especially in the regions of repeated nucleotide sequence [20]. Thus, this advancement came at the cost of shorter read-lengths and decreased accuracy compared to Sanger sequencing. Consequently, the computationally intensive optimization efforts following the introduction of sequencing technology led to increases in its accuracy along with a decrease in price [21]. This resulted in a drastic transformation of the field of genomics and other areas of biology with new applications of sequencing technology still being devised today [22,23].

This technological advancement in the analysis of the genome at a large scale led to its application on cancer via two major projects, one beginning in 2001 and another in 2005 [24] (see https://cancergenome.nih.gov/abouttcga/ overview/history). The first project, called the Cancer Genome Project, was based in Wellcome Trust Sanger Institute. The project focused on the cataloging of recurrently detected somatic mutations that are found by comparing the tumor tissue of a patient to a normal tissue in the same person. Researchers studying the mutations were able to differentiate them into the

"driver" mutations, which impair specific genes to propel the development of tumor, and the "passenger" mutations that occur randomly with inconsequential effect on the tumor growth. These efforts resulted in a large online database of somatic mutations in cancer called Catalogue Of Somatic Mutations In Cancer (COSMIC) [25]. The other major project in sequencing cancer tissues was the Cancer Genome Atlas (TCGA) project in 2005. The purpose of this project was to better understand the genetic mechanisms involved in cancer so that it can be better diagnosed and treated (see https://www.genome.gov/10000905/national-advisory-council-for-human-genome-research/). Unlike the study of normal tissues, cancer analysis using the sequencing technology can involve further complications. For instance, when a bulk tissue of a tumor is excised, tumor microenvironment within the tissue often also includes non-cancerous cells [26]. Aran illustrates how this decrease in the tumor purity can have significant effects on the analysis of cancer. In addition to the tumor purity, regional sequencing studies have shown that genetic intratumoral heterogeneity makes it difficult to capture all of the genomic landscape of a cancer tissue [27]. Despite these limitations, the sequencing technologies truly revolutionized the field of cancer genomics.

In order to understand how genomic information is obtained from the sequencing data, several terms need to be defined. Coverage in sequencing refers to the number of times the genomic region is "covered" or sequenced, assuming a constant length of reads mapping in a random distribution [28]. Once coverage of the sequence is determined, allele frequency (AF) of a single nucleotide variant (SNV) can be calculated. This allele frequency, or allele ratio, of a SNV can be calculated by dividing the number of reads supporting the reference nucleotide letter by the number of reads supporting the other

nucleotide letter at that particular position [29]. There are several bioinformatics tools that both identify all SNVs and calculate the AF for them. These often filter out the reads that are uncertain in their calls or mapping before identifying all SNVs and use a statistical model to classify the SNVs [30]. Copy number variations (CNVs) are defined as genomic regions with a copy number ratio that deviates away from the normal ratio of 2 and are generally identified by comparing the relative copy number ratio calculated from the read coverage across the genome [31].

The ability to analyze multiple samples using the sequencing technology at a relatively cheap price opened up new opportunities for various cancer genome-wide association studies. Comparing cancer cases at such a large-scale highlighted specific genes most frequently mutated in particular subtypes of cancer, which subsequently provided clinical benefits such as the identification of specific targets for different cancer types. According to one TCGA study on colorectal cancer, eight most commonly mutated genes in non-hypermutated CRC cases were *APC*, *TP53*, *KRAS*, *PIK3CA*, *FBXW7*, *SMAD4*, *TCF7L2,* and *NRAS*, while the most commonly mutated genes in hypermutated CRC cases were identified as *ACVR2A*, *APC*, *TGFBR2*, *MSH3*, *MSH6*, *SLC9A9* and *TCF7L2* [32].

In addition to determining genes with the most recurrent mutations, identifying somatic mutations using sequencing technology allowed for the analysis of the specific mutational mechanisms. In 2013, as part of the Cancer Genome Project, researchers classified each substitution mutation found in a cancer sample into one of the 96 possible tri-nucleotide sequences organized by six different base substitution classes and examined the relative distribution [33]. The researchers showed that by applying this

holistic view of the landscape of the somatic mutations in each cancer sample across many cancer types, they observed and validated distinct patterns or signatures. Many of these signatures were assigned specific functions or characteristics that were most probable based on the association study. This analysis offers a new way to better understand the mutational mechanisms from somatic substitution mutations.

Modes of Evolution and Sequencing of CRC

Since the introduction of next generation sequencing technology, improved ability to accurately characterize cancer genomes allowed researchers to explore and challenge the idea of sequential progression in carcinogenesis of CRC and this effort resulted in three additional evolutionary models to address carcinogenesis. One recent study on the distribution of somatic mutations in CRC proposed an extension of the linear model into a quick eruptive accumulation of mutations in the polyp followed by subclonal competition and a plateau of extensive mutation accumulation in the resulting cancer [34]. This phenomenon results from a lack of selection pressure in combination with high rate of mutation accumulation. One noteworthy aspect of this model, also known as the 'Big Bang,' is that it describes how early mutations shape the high intratumoral heterogeneity (ITH) observed in the late stage of CRC and may be associated with a more aggressive clinical behavior and decreased rates of survival [27].

Another study postulated a parallel evolution involving a separate lineage of private, cancer- and adenoma-specific mutations branching out from the early clonal mutations shared between the two tissues [35]. Due to this divergence, cancer exhibits a different mutational architecture than the traditionally expected expansion of the subclonal populations present in the polyp. In parallel evolution, driver mutations may be

present in subclones and multiple independent subclonal expansions may persist given that they may be of equal fitness. Thus one subclone does not confer a selective advantage over another subclone, which may in fact have the same driver mutation but unique private mutations. In some cases, the polyp accumulates a greater number of mutations than cancer, while in others the cancer accumulates more mutations than the polyp, suggesting that the number of mutations alone cannot determine whether a CRC will undergo parallel evolution.

Lastly, the possibility of cancer evolution following a simple neutral power-law was explored based on the finding that some cancers exhibit several distinct distributions of the allele frequency of somatic mutations in their cancer lineage in which one evinces selection pressure while another does not. The notion is that the distribution of passenger mutations with low allele frequencies with respect to the clonal mutations near the allele frequency of 0.5 would follow a 1/f distribution [36]. On the other hand, a lineage exerting selection pressure would have driver mutations represented as an additional subclonal peak located between the peak of the passenger mutations and that of the clonal mutations [37].

These four Modes of Evolution (MOEs), stepwise, eruptive, parallel, and neutral, offer possible explanations for the temporal relationship among the different types of genome wide alterations as well as intra- and inter-tumoral heterogeneity. The importance of knowing the features of MOEs with the highest impact on the polyp or tumor's behavior is highlighted by the genetic evolution in glioblastoma multiforme (GBM). In one study, the degree of persistent mutations from the primary tumor that were also present in the recurrent tumor differed based on the MOE of the primary tumor

[38]. The authors found that the recurrent tumors carried 75% of the mutations present in the primary tumor with linear MOE compared to those with parallel MOE in which recurrent tumors shared only 25% of the mutations in the primary tumor. The recurrence of the cancer, which rarely can be successfully treated and cured, was genetically represented by those early mutations present in the primary tumor that had developed resistance to chemotherapy. If it were possible to recognize all functionally relevant mutations in the primary tumor, and develop treatments for each of these mutations, conceivably it would be possible to prevent recurrences. Thus, studying the MOEs and understanding the characteristics of clinical cases in relation to the evolutionary path that led to malignant transformation may be leveraged to improve accurate prognostication and provide targets for personalized treatment options.

We previously reported that at least 10% of CRC have the contiguous residual polyp of origin (CRC RPO+) identifiable in the surgically resected specimen [39]. We performed whole genome sequencing and analysis of such CRC RPO+ cases, which included matched peripheral blood leukocytes, normal colon a minimum of 8 cm distant from the polyp and/or cancer edge, the cancer adjacent polyp (CAP) and the contiguous CRC (**Fig. 1**). Comparative analysis between CRC RPO+ and CRC without residual polyp of origin (RPO-) revealed essentially the same histology, gene expression patterns, mutated genes/pathways, as well as the same stage-adjusted disease free and overall survival. Similarly, the CAP component is highly likely to represent the intermediary state between normal colon epithelium and CRC RPO+ because it is in direct contiguity with the cancer. This strongly argues that CRC RPO+ is a valid model to study neoplastic transformation in the colon [39].

In this study, we analyzed patterns of somatic mutations in the CRC RPO+ cases to determine the relationship between different MOEs in the transformation from normal colon cells to CRC and mutational characteristics of CRCs, cancer adjacent polyps, and cancer free polyps. To determine this relationship, we analyzed the same whole genome sequence data from our previous study [39]. The cases included corresponding distant normal colon epithelium, CAP and the cancer from 13 CRC RPO+ cases and 14 CFP cases with matching distant normal or PBL for each case, which make up more than 80 tissue samples extracted from 27 patients. Cases without the matched normal colon epithelial tissue were excluded. 11 of these cases were clinically determined to be aggressive, having either recurred or presented as advanced stage IV disease.

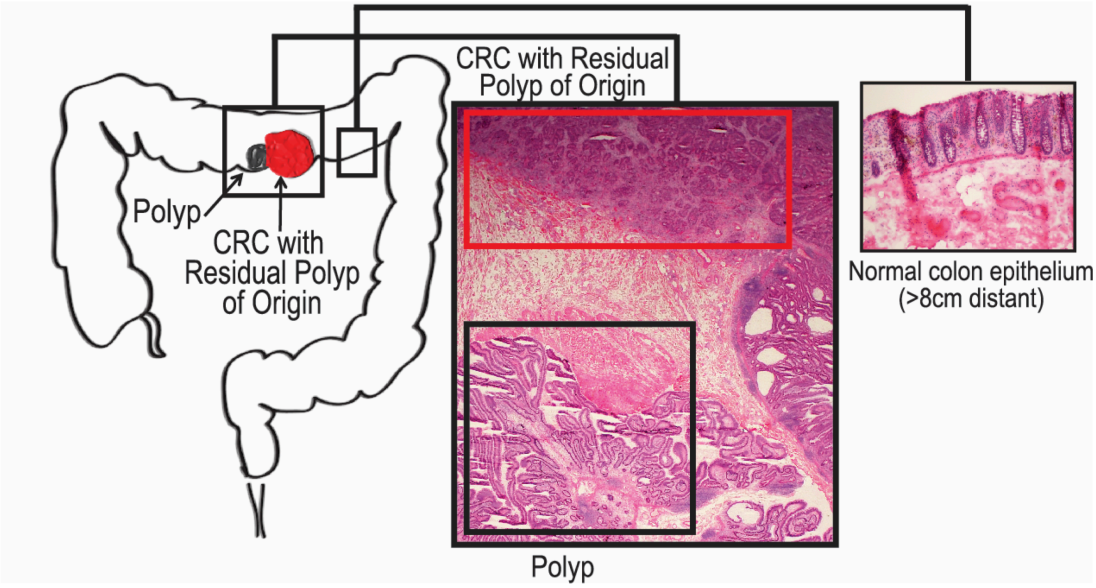


Figure 1: Depiction of the regions of biopsy
CRC with residual polyp of origin (top red box) refers to cancer that has a polyp located physically adjacent to the cancer (bottom black box). Matching normal colon tissue was also harvested at a distance > 8cm from the cancer.

MATERIALS AND METHODS

Patient sample characteristics

All tissues were collected from patients consented to the IRB approved Biobank for Gastrointestinal Health Research [BGHR] (IRB 622-00, PI LA Boardman) at Mayo Clinic between 2000-2016. 1 cm^2 portions of surgically or endoscopically resected cancers from patients with CRC RPO+ were snap frozen in liquid nitrogen and maintained long term at -80 °C. Cancer adjacent polyps (CAPs) were identical to cancer free polyps (CFPs) in size (1 to 2cm, 2-5 cm and >5cm) and histology (tubular or villous subtype), and the degree of dysplasia (low-grade). Matched normal colonic epithelium were collected at least 8cm away from the polyp/tumor margin. This study did not include subjects with family history of FAP or Lynch syndrome and any other hereditary CRC or inflammatory bowel disease.

Tissue preparation and Whole Genome Sequencing

An Hematoxylin and Eosin slide circled by a pathologist to enrich for distant normal colon epithelium a minimum of 8 cm away from the polyp or tumor edge, polyp and cancer tissues was used a guide slide for macrodissection of these three tissue compartments. DNA from peripheral blood leukocytes (PBL) from the patients was obtained on a subset of these patients. DNA was extracted using the PureGene method and was quantified with appropriate kits on the Qubit Fluorometer. Samples were sequenced at the Broad Institute on the Illumina HiSeq X instruments producing 150 base pair, paired-end reads to meet a goal of 30x mean coverage. All data from a particular sample was aggregated into a single BAM file using the Picard Tools (https://broadinstitute.github.io/picard/).

Mutation Frequency Detection

Four different somatic variant callers were used to identify SNVs in the polyp and cancer against the matched normal tissue or PBL with default options: MuTect, SomaticSniper, Strelka, and VarScan [30,40–42]. We only took SNVs detected by at least two callers. Variant allele frequencies for those SNVs were calculated from sample BAM files for each patient using an in-house script. For functional annotations of the variants, we used Variant Effect Predictor (http://www.ensembl.org/Tools/VEP).

Mutational Motifs Calling

From the list of SNVs called in cancer, somatic SNVs that were also called in polyp were subtracted from the list to ensure mutual exclusivity between the cancer and polyp SNVs. Cancer sample in the cases A04, A09, A11, A14, and A15 did not have enough somatic SNVs to exclude the SNVs found in polyp and thus all somatic mutations found in cancer were included. Each somatic SNV within a sample was categorized into the corresponding transversion or transition substitution mutation into one of the 96 tri-nucleotide possibilities. After normalization, correlation coefficient was calculated based on these vectors of 96 integers for each sample and UPGMA-based hierarchical clustering was performed on the samples based on this coefficient. The intensity of the colors was adjusted within each of the six panels per sample for easier detection of the patterns. All the statistical analyses were performed using R software. Heatmaps were generated using the ggplot() function in the R package, ggplot2. Hierarchical clustering was performed with the hclust() function with default parameters. Correlations coefficients were calculated using the cor() function with the Pearson method.

Pathway from Related Genes

For each of the CFP, CAP, and cancer, a list of genes with somatic mutations was submitted to the Database for Annotation, Visualization and Integrated Discovery (DAVID) for functional pathway analysis. To account for the background mutation rate, only the genes determined by the MutSig to be significantly mutated were used in the analysis (p-value <0.05). The number of significantly mutated genes for CFP, CAP, and cancer were determined to be 195, 123, and 137 respectively. The results of the multiple hypothesis tests were corrected using Benjamini method (FDR < 0.05).

CNA Analysis

The regions of deletion and duplication was genotyped using CNVnator using a bin size of 200bp and only the regions with a copy number of > 1.75 and < 2.25 in normal samples were considered. To further filter the CNAs, only the regions with copy number difference greater than 0.2 with respect to normal tissue were chosen. A pairwise similarity metric, M, is based on Jaccard similarity coefficient was defined as follows:

$$M = \sum_{per\ Chr} \frac{\sum R_{dup,dup} + \sum R_{del,del}}{\sum {R^1}_{dup} + \sum {R^1}_{del} + \sum {R^2}_{dup} + \sum {R^2}_{del} - (\sum R_{dup,dup} + \sum R_{del,del})}$$

The similarity metric M is equal to the region, R, of duplications or deletions in both samples over all regions of duplication or deletion for each chromosome. Because these are scores per chromosome, calculating the summation of these scores represents a similarity score between a pair of samples across their entire genome with the exception of X and Y to ensure comparison across genders. For each tissue type, chromosomes with significantly recurrent aneuploidies compared to others were determined by a Wilcoxon signed-rank test. With a pairwise similarity metric across all the samples, UPGMA-based hierarchical clustering was performed. Heatmaps were generated by dividing the genomic

regions into segments of 50kb and using the ggplot() function in the R package, ggplot2. Hierarchical clustering was performed with the hclust() function with default parameters.

RESULTS

Utility of mutation AFs across neoplastic transformation for MOE classification

We observed distinct patterns in the allele frequency distribution of each pair of cancer and matching polyp characterized by 1) two gaps in AF; 2) one gap in AF; or 3) no gap (**Appendix Figure 1-14).** We hypothesized that these patterns are representative of CRC evolution. Consequently, we derived an expected pattern of AF distributions in CRC and its corresponding polyp for each MOE based on its key characteristics. An illustration of the clonal lineages corresponding to each model is shown in **Figure 2**. For stepwise and eruptive MOEs, it is expected that most SNVs are shared between the adenoma and its corresponding cancer. Moreover, selective pressure is a common feature in stepwise MOE and can increase the AF of certain subclonal mutations nearly to the level of AF present in shared clonal mutations. For parallel MOE, where the adenoma and cancer branch early in their evolution to independent pathways, most SNVs are expected to accumulate after the branch point between a polyp and cancer. Thus, most of these SNVs are polyp or cancer specific, rather than shared. One important distinction between stepwise and parallel MOEs is that the parallel MOE has adenoma and cancer separately evolving along their respective subclonal lineage, implying distinct adenoma- and cancer-specific mutations conferring a growth advantage for each tissue compartment. Thus, the AF of private mutations in polyp compartment of parallel MOE cases would both be close to 0.5 (blue dots in **Fig. 2**), while this is not the case for stepwise MOE. A key feature of neutral MOE is the little to no selective pressure

represented by a lack of clear distinction between early-shared (gray cluster) and late-shared mutations (the rest). CRC originating via an eruptive MOE is characterized by the development of all shared clonal mutations prior to the pre-cancer polyp phase followed by little to no selection so that the shared clonal mutations early in transformation would include clusters of early mutations (brown, green, and light blue cluster) in addition to the grey cluster of SNVs.

The anticipated AF distribution for the union of somatic SNVs per patient (i.e., discovered both in adenoma and cancer) is a characteristic feature of MOE in both the adenoma and cancer compartments. Gradual transition from adenoma to cancer in stepwise MOE will result in a gap in the AF distribution in cancer and small or no gap in adenoma. However, no such gaps are expected in the distributions for neutral MOE, as SNVs with all AFs are shared between adenoma and cancer. Long independent evolution of adenoma and cancer in parallel MOE will results in clear gaps in the two distributions. Early shaping of the shared clonal mutations in eruptive MOE will results in a gap in AF distribution of adenoma, but at no or less pronounced gap in AF distribution of cancer.

Purity of a tissue, i.e., fraction of malignant cells, is generally determined by a pathologist visually inspecting histological slides of the tissue and the purity level varies from sample to sample [26]. Lower purity level in adenoma sample would shift down the overall somatic SNV AF distribution in adenoma towards zero and could potentially decrease the gap observable in the 2D plot. However, while purity level can shift, shrink, or expand the AF distribution and affect the absolute AF, the relative pattern of the distribution as a whole remains the same. Thus, comparison of the relative AF between

the shared and the private mutations can be used to infer the existence of selective pressure in a CRC without the influence of the purity level difference across samples.

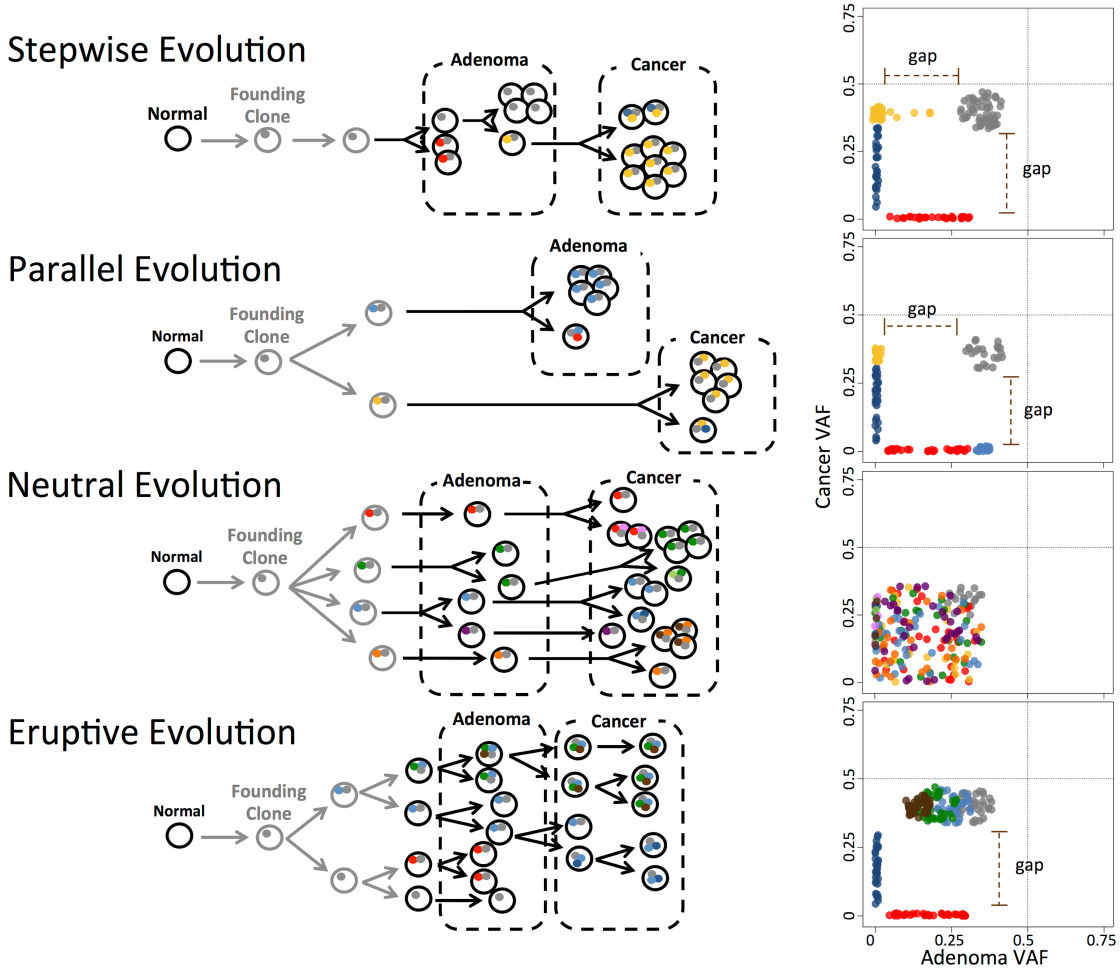


Figure 2: Schematic representation of the four Modes of Evolution (MOEs) in the transformation from adenoma to colorectal cancer
At the very top is the classical sequential stepwise progression. Next, the parallel model depicts the branching of the subclones. Neutral evolution has subclonal expansion in the absence of selective pressure. Lastly, the eruptive 'Big Bang' model represents the early bursts of genomic disruption. Circles and dots depict cells and mutations respectively. Colors correspond to different clones/subclones. Plots on the right represent 2D AF distributions of somatic variants in polyps and cancer with dotted lines on 0.5 AF. These distributions are characteristic of each mode and can be used for MOE prediction.

CRCs also exhibit large chromosomal aneuploidies with deletion or duplication of the entire chromosomes and/or chromosomal arms, which can be characterized in a MOE-specific manner. Aneuploidies are expected to be noticeable beginning in adenoma

for eruptive MOEs, in which most of the copy number alterations happen early in the lineage and the cancer grows out of a clone present in the polyp stage. In an independent evolution of polyp and CRC, as in parallel MOE, aneuploidies should only be observed in cancer. Similarly, in the stepwise evolution, deleterious early copy number alterations would be selected out, leaving only the chromosomally stable clone in polyp stage to grow into cancer, at which the growth could tolerate larger scale chromosomal changes. The neutral MOE's copy number alterations status can be more difficult to interpret because the lack of selection could imply a possibility of smaller copy number alterations but never a larger, more damaging copy number alteration that only cells with selective advantage could tolerate.

Here, it is crucial to note that aneuploidies would not affect the shape of the distribution in the 2D plot for several reasons. First, large regions of the genome are chromosomally stable in most of the polyp cases. In fact, only three out of the 13 CAP cases have greater than 10% of its genome affected by aneuploidies. Another reason is that even if the adenoma is affected by aneuploidies, they are mostly subclonal and will not shift the AF of the private mutations significantly. This, along with the simultaneous shift in AF between the shared and the private SNVs mentioned above, support the notion that aneuploidies are unlikely to affect our interpretation of the 2D plot in classifying cases by MOEs. All of these reasons also apply to genome doubling. Genome doubling, which is found to be commonly associated with increased rate of evolutionary growth in colorectal cancer, occurs early in the development and can either affect both adenoma and cancer to raise the AF distributions in both axes or, again, affect the entire tissue so that the AF in both the private and shared mutations shift together. Therefore, using the

relative position of the AF in private mutations with respect to the shared mutations serves as a strong criterion in determining the MOE of a case.

Comparative example of stepwise and eruptive MOEs

Let's consider two cases of neoplastic transformation to CRC in our cohort: in case A03, where we classified MOE as stepwise, and in case A09, where we classified MOE as eruptive (**Fig. 3**). In each patient, two stages of adenoma (tubular and villous) were observed, harvested, and sequenced. In these cases, the tubular and villous polyp along with the corresponding resultant cancer were in direct contiguity and present on the same histology slide, which likely represents the tissue compartments involved in the malignancy that patient A03 developed. Tubular adenoma in A03 had 5,611 somatic SNVs with no aneuploidies. Tubular adenoma in A09 had 8,786 somatic SNVs with apparent aneuploidies. As expected from the classical Fearon and Vogelstein model, both the tubular and villous polyps in A03 had a stop mutation in *APC*, which is a gene recognized to be mutated in many CRC and considered to be involved in initiating neoplastic transformation in the colon. Each of the patients also had mutations in one of two well-known cancer-driver genes: in *KRAS* and *TP53* in the patient case A03 and in *TP53* in the patient case A09. The introduction of the *TP53* also correlated with observation of large amounts of aneuploidies (**Fig. 3B,D**). In both patient cases, evolution to villous adenoma did not change the copy number profile though the copy number alterations became more pronounced, likely as a result of better sample purity in the villous compared to the corresponding tubular polyp compartment. This is particularly noticeable in case A03, in which more somatic SNVs are detected in the villous polyp than in its corresponding tubular adenoma, 14,393 vs. 5,611. Consistently, AF of early

mutations originated in tubular, including stop mutations in *APC* and *KRAS*, are increasing and are centered close to 50%, suggesting that one clone dominates this stage. In A09, count of SNVs increase only slightly, 8,893 vs. 8,786, with AF of early mutations, including stop mutations in *APC* and *TP53*, unchanged. As an average AF of early mutations is significantly below 50%, mutation-containing cells likely constitute only a small fraction of all cells in the villous polyp.

The copy number profile in the cancer from patient A09 is the same as, but much more pronounced, in the villous tissues, suggesting higher purity of cancer cells in the cancer compartment than in the villous tissues. In agreement with this, more somatic SNVs are detected in cancer compared to the villous polyp, 11,357 vs. 8,893, and the AF of early mutations in the adenoma are centered close to 50%, suggesting high purity level in this sample. Deletion of the other (not mutated) copy of *TP53* gene in the villous compartment leads to an even higher AF of a stop mutation in this gene. New mutations only found in cancer have much lower AF, indicating that these mutations represent subclones in the cancer. Large aneuploidies becoming progressively more definitive since their introduction early in the lineage is consistent with eruptive progression.

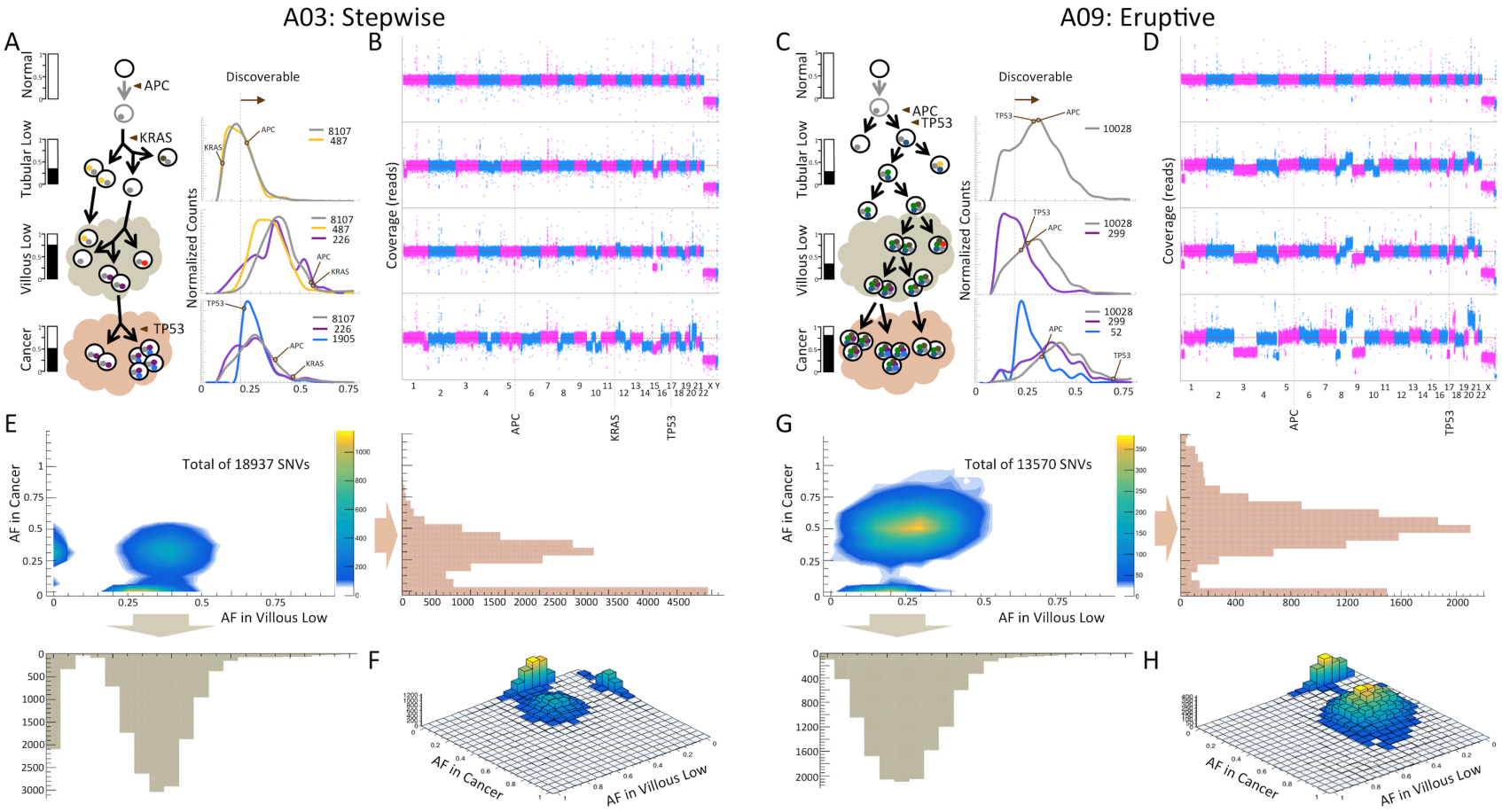


Figure 3: Example of a stepwise (A03) and an eruptive MOE (A09) revealed by somatic mutations analysis
A&C) Schematic representation of models for origin, presence, and propagation of clonal and subclonal mutations at each stage of transformation. Mutations in *APC*, *KRAS*, and *TP53* are labeled at the corresponding stages of the evolution. The bars on the left represent sample purity from SNV AF and CNA analyses. A) Distributions on the right represent the mutations shared by tubular and villous low adenomas, and cancer in gray, those shared between tubular and villous low adenomas in yellow, those shared between the villous low and cancer in purple, and those that are cancer-specific in blue. C) Distributions on the right represent the mutations shared

by tubular and villous low adenomas, and cancer in gray, green, navy blue, and brown colors. Those shared between the villous low and cancer are represented in purple and cancer-specific in blue. The distribution on the right has these four colors represented in gray for simplicity. B&D) Genome copy number profiles at each stage. In case A03, large aneuploidies are observed only in the cancer stage. In case A09, large aneuploidies are observed in the tubular stage and are maintained until the cancer stage. E&G) Distributions of AF for somatic SNVs are consistent with stepwise MOE in the case A03 and eruptive MOE in the case A09 (**Fig. 2**). F&H) 3D representation of the AF distributions of SNVs for each case. The height of the distributions, which shows the number of mutations, can be used in the comparison of the shared SNVs and private SNVs during the MOE classification.

Contrary to this, aneuploidies and CNAs in A03 are present only in cancer, in which a stop mutation is observed in *TP53*. However, number of detectable somatic SNVs decreases as compared to villous, 8,831 vs. 14,393, with AF of mutation observed in adenoma also decreasing. This is most likely due to lower purity in the cancer component. New mutations found only in cancer are centered at a similar AF despite being subclonal, suggesting a significant selective advantage. All these observations are consistent with the gradual accumulation of mutations followed by a selective pressure that progressively alters the genomic landscape mostly with SNVs until the late stage of cancer, at which point there are large aneuploidies and CNAs.

Rules and for classifying MOEs

We defined a set of rules suggesting MOE based on each characteristic signature that can be observed in spatial-temporal pattern of SNVs (**Table 1**). Five rules were defined to describe characteristics of each MOE. These rules are related to the number of SNVs, the shape of the SNV AF distribution in adenoma and cancer, and the progression of CNAs. These rules were based on the theoretical expectations for the four considered MOEs (**Fig. 2**).

Rule #1 is comparing the number of adenoma- and cancer-specific somatic SNVs with the number of somatic SNVs shared between them. Rule #2 is testing whether a gap exists in the AF distribution of somatic SNVs in a 2D plot of adenoma vs. cancer (**Fig. 2**). Since selection pressure within a lineage would shift the private SNV AFs further away from the shared SNV AFs, a larger gap would signify mutational architecture in later stages that is far different from its shared, early clonal mutations. Rule #3 compares the spatial distribution of the shared SNV AFs with respect to the private SNV AFs. In

evolutionary scenarios with selective pressure, almost all of the later cell population would derive from a cell that acquired driver mutations early in the lineage. The high prevalence of this cell leads to a private SNV AF that nearly matches the SNV AF of the shared, clonal SNV AF. Rule #4 examines whether the early shared SNVs and the later shared SNVs are distinguishable based on AF distribution. Although similar to rule #2, rule #4 requires both adenoma- and cancer-specific mutations, which will filter out neutral evolution in which the lack of selection pressure leads all the mutations in the adenoma to be carried forward to the cancer. Similarly, rule #4 filters out eruptive evolution given that in eruptive evolution the majority of the mutations found in the cancer similar to those in the adenoma due to the late steady state in which mutations infrequently accumulate in the CRC beyond those that rapidly accumulated in the early precancerous polyp phase and persisted into the cancer. Lastly, rule #5 compares the earliest time point at which large aneuploidies are noticeable.

We then applied majority vote from all rules to classify MOE for each analyzed CRC case (**Table 2**). We found that approximately 70% of CRCs in our dataset evolved in a stepwise or a parallel MOE. One case demonstrated an eruptive MOE and three cases neutral MOE. Cases A11 and A13 both had a gap in the SNV AF distribution in the adenoma but not in the cancer, which corresponded to none of the characteristics of MOEs outlined earlier. For A13, the majority of the rules did not apply because of the lack of distinction between the private and the shared SNV AF distribution in addition to the unusual characteristics in the SNV AF gap between the adenoma and its related cancer. One explanation for the unusual gap in the 2D plot is a low purity level in the cancer compared to the adenoma, which is supported by the overall low AF distribution

in cancer near 0.25. However, if that were indeed the case, it would still not explain the fact that the private mutations in cancer appear to have a higher AF than the shared SNVs. In fact, pathology review indicates that the macrodissected portion of the cancer had 60% tumor density (**App. Table 1**). For these reasons, cancer-specific mutations shifting relative to the mutations shared between both adenoma and cancer can be attributed to another possible explanation that the CRC and the physically adjacent CAP for cases A11 and A13 arose from different clones independently rather than the CRC growing from the same clone as the CAP. Lastly, there is also a possibility that these cases represent variations of the four MOEs our rule sets fail to capture or that these cases may represent an additional MOE such as an extremely rare polyclonal [43].

Table 1: Rules to classify MOE for each case of neoplastic transformation

Rules \ MOE	Stepwise	Parallel	Neutral	Eruptive
#1: Count of private SNVs relative to shared SNVs	Large fraction of private in cancer and in polyp as well as shared	Majority is private in cancer and in polyp	Majority is shared	Majority is private in polyp and shared
#2: Gap in the SNV AF distribution in adenoma and in cancer	Yes/Yes	Yes/Yes	No/No	No/Yes
#3: Position of private SNV clusters relative to the shared SNV cluster	Adenoma-specific SNV AF is lower than the shared SNV AF	Both adenoma and cancer-specific SNV AFs are equal to the shared SNV AF	No private mutations	Both adenoma- and cancer-specific SNV AFs are lower than the shared SNV AF
#4: In-distinguishable early shared from later shared SNVs	No	No	Yes	Yes
#5: Numerous Aneuploidies start in	Cancer (Late)	Cancer (Late)	N/A	Adenoma (Early)

Table 2: Rule-based classification of each neoplastic transformation case
The label s stands for stepwise, p stands for parallel, n stands for neutral, e stands for eruptive, and s,p for a combination of stepwise and parallel. For each rule, the samples were given the most likely MOE assignments. Then, each sample was assigned the final MOE with the majority count across the five rules.

Cases\Rules	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5	Conclusion
A02	s	s,p	p	s,p	s,p,n	s,p
A03	s	s,p	s	s,p	s,p,n	s
A04	n	n	n	n,e	n,e	n
A07	s	s,p	p	s,p	s,p,n	s,p
A08	p	s,p	p	s,p	s,p,n	p
A09	e	e	s,e	n,e	n,e	e
A10	n	n	n	n,e	-	n
A11	s	-	s	s,p	n,e	s
A12	s	s,p	p	s,p	-	s,p
A13	p	-	-	-	n,e or s,p,n	most likely p
A14	n	n	n	n,e	-	n
A15	e	s,p	p	s,p	-	p
A16	s	s,p	s	s,p	-	s

Thus, we only applied the rules #1 and #5 that were pertinent to the case A13 and this limitation led to a classification that is less reliable than other cases but the relevant rule-based characteristics strongly suggested parallel MOE. It should be noted here that due to the nature of synchronous residual polyp of origin, all patient cases exhibit the property of branching evolution to some extent, even though it is assumed that the cancer sample developed from the adenoma. This is the reason we labeled a few cases to have an MOE of s,p as we could not conclusively classify these cases as having stepwise vs. parallel MOE. Also, the striking feature of neutral MOE is its lack of selective pressure in its lineage, which allows aneuploidies to begin at any point in time. All samples with observable aneuploidies were given the additional neutral MOE assignment for rule #5

because of this inability to predict the initiation of aneuploidies. Lastly, samples without significant aneuploidies were not considered for rule #5.

Mutational signatures and MOEs

Next, we broke down the somatic mutations into 96 tri-nucleotide motifs for each sample, as Alexandrov previously did in cancer in order to decipher mutational signatures [33]. Then, we performed a pairwise comparison, and ordered them by hierarchical clustering in order to identify any mutational patterns specific to different MOEs (**Fig. 4**). C>T transitions and C>A transversions seem to contribute the most in distinguishing the clusters apart. Additionally, the similarity in the mutational patterns of cancer and their matched adenoma samples imply that the mechanism resulting in the shaping of the mutational spectra is determined even before adenoma and remains relatively stable in the corresponding cancer.

Clustering of the CAP and cancer samples based on their somatic mutational spectra clearly show three major groups of the samples with the same subtypes: case A16 being in one, cases A10 and A12 being in another, and the rest being in another group. In other words, the subtyping of microsatellite-high (MSI-H) or chromosomal instability (CIN) case contributes more significantly to the substitution mutation patterns than the MOE assignment does. Nevertheless, the same subtype does not guarantee the same MOE assignment and this is consistent with two microsatellite-high (MSI-H) cases of A10 and A12 having a different MOE. Although the same underlying genetic hypermutability results in the two cases having a similar number of mutations, A10 cancer had a neutral MOE while A12 cancer underwent a stepwise or parallel MOE. The genetic hypermutability in both A10 and A12 cases were confirmed in their mutational

patterns that are similar to signature 6. According to Alexandrov, signature 6 is commonly observed in CRC and is presumed to be associated with compromised DNA mismatch repair (**App. Fig. 12&13**).

Additionally, the cluster analysis illustrates that the MOE classification is not solely dependent on the number of mutations or the MSI status. Despite the A16 cancer sample having approximately 10K somatic mutations, it is classified as stepwise MOE as most of the other cases. Case A16, due to its unusually high mutational burden that is approximately 62 times the amount of somatic mutations found in CIN cases and 6.7 times the amount found in MSI-H cases, was presumed to have mutations in POLE. This was confirmed when the adenoma and cancer mutational spectra were compared to signature 10 in the COSMIC database (**App. Fig. 14**). Since signature 10 is commonly observed in CRC and is speculated to be associated with variants in DNA polymerase epsilon, this suggested a defective POLE gene [33]. A missense somatic mutation call in POLE gene with an AF of 0.25 in the cancer sample of the case A16 accounted for only a part of the story since the mutational spectra tell us that the mutations in POLE gene happened early in the evolution for such patterns to form both in adenoma and in cancer. Indeed, we found two missense mutations in the matched normal sample of the case A16. While these could be either germline mutations present throughout or somatic mutations that occurred early in development, these mutations could possibly explain the ultrahigh-mutation rate observed in both the polyp and cancer. The first missense mutation is a known SNP on chromosome 12 at position 133220526 that is predicted to have a deleterious effect with a SIFT score of 0.03, which is contrary to a benign effect predicted by the PolyPhen2 score of 0.104 (SNP ID rs5744934). The second missense

mutation, which we believe to be responsible for ultrahigh -mutation effect, is a variant that has not been previously reported and is located on chromosome 12 at position 133220556. Both SIFT and PolyPhen2 predict the resultant amino acid change from arginine to proline to have a damaging effect with scores of 0.01 and 0.997, respectively. These results imply that defects in specific pathways occurring early in the lineage might contribute more to the shaping of these mutational spectra than the particular MOE that led to the tumor, though the relationship or cause versus effect of the ultrahigh mutation rate and the MOE requires further study in a larger sample set.

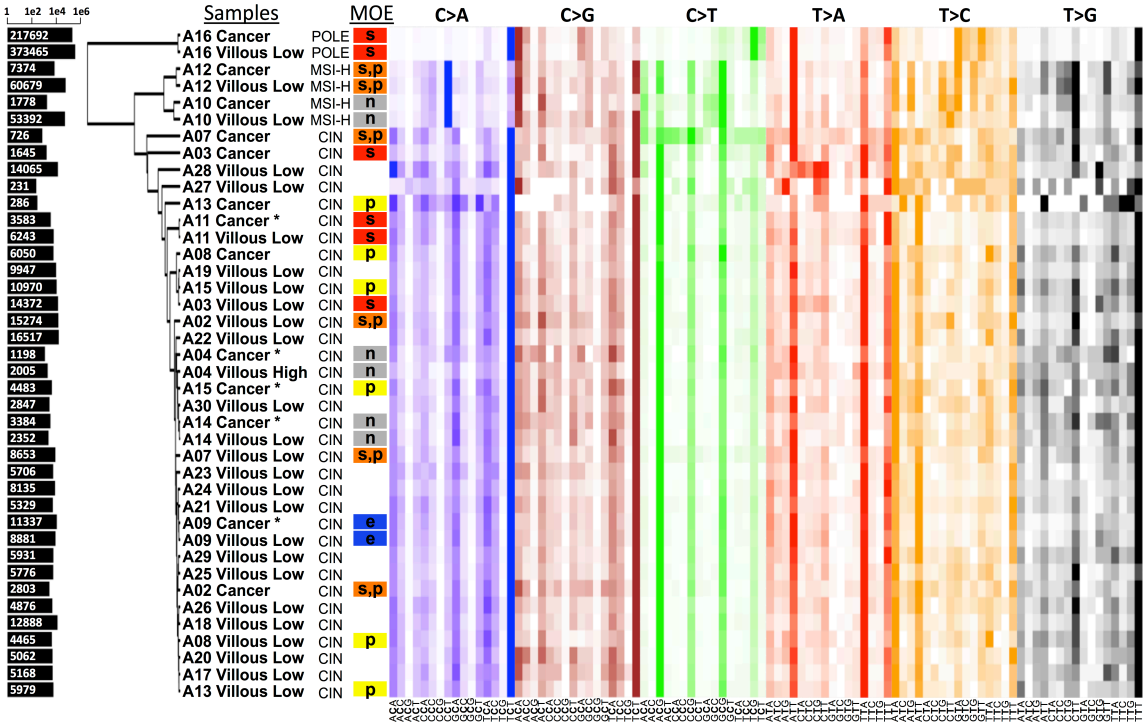


Figure 4: Heatmap of the mutational signature analysis and hierarchical clustering of all cases

Each colored panel represent the 16 possible tri-nucleotide combinations corresponding to the respective transversion or transition type. The color intensity indicates the proportion of the particular mutational signature for that substitution mutation. Bars on the left represent the total number of SNVs for each sample in log scale. Clearly visible linear pattern over all the samples with high color intensity suggests mutational components that are similar in the majority of cases. The samples are clustered into three major clusters with A16 samples being in one, A12 and A10 in another, and the rest in one cluster. A few samples, indicated by the asterisk, had too few cancer-specific

mutations. For these samples, the somatic mutations included those found in the villous low adenoma stage.

Enrichment in pathways

With the list of genes that are significantly mutated in CFP, CAP, and cancer cases, compromised functional pathways important in the shaping of the tumor can be identified. Only feature represented were the various regions of repeating sequence across both adenomas and cancer. For CFP cases, HTLV-1 infection also had significant number of mutated genes in its pathway. Other less significantly affected pathway categories include glycoproteins, disulfide bonds, and secretory.

Aneuploidies and MOE

To determine if a pattern in the CNAs have specific connections to the MOEs, we devised a pairwise similarity metric characterizing a chromosomal region of duplications or deletions present in both of the samples. The scoring emphasizes only the similarity in the pairs and thus gives a small regional variation the same weight as an entire chromosomal aberration as long as they are present in both samples. A chromosome can have a score between 0 and 1, and the score is higher if more samples have overlapping regions of copy number aberrations for this particular chromosome compared to the other chromosomes. Because these are scores per chromosome, calculating the summation of these scores represents a similarity score between a pair of samples across their entire genome. Samples corresponding to each CFP, CAP, and cancer tissue types were separately compared and clustered based on these values (**Fig. 5**).

In addition to comparing the CNAs across the samples, this similarity metric also allowed per-chromosome analysis to identify chromosomes with more recurrent CNAs compared to other chromosomes for each of the CFP, CAP, and cancer tissue type. The

table below the three panels indicates the chromosomes with this information. Because we often observe higher aneuploidy level in cancer as opposed to adenoma, a comparison across tissue types may not be possible. Nevertheless, identification of potential markers is an essential step towards understanding the clinical significance of tissue types in addition to the MOE classification. Chromosomes 7, 17, 20, and 18 had the most recurrent copy number alterations across cancer cases, while chromosomes 1, 16, 17, 18, 15, 20, and 7 were most recurrent across CAP cases. For CFP cases, chromosomes 21, 22, 1, 13, and 20 had the most recurrent copy number alterations.

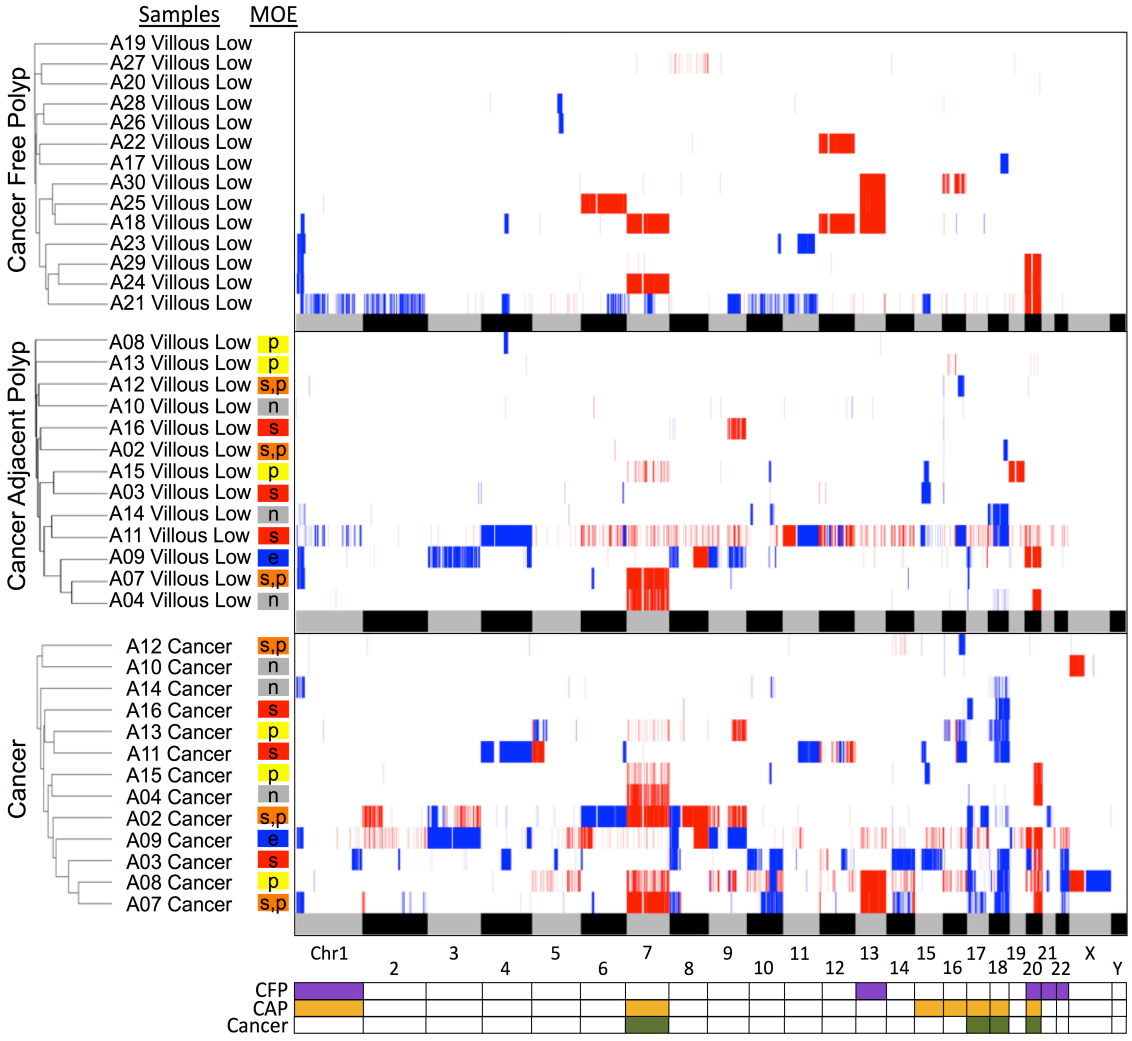


Figure 5: Heatmap of the DNA copy number analysis and hierarchical clustering of the cases by the tissue type

CNAs of the entire genome in each sample are indicated by either deletions (blue) or duplications (red). The alternating grey and black bars at the bottom of each panel represent the spanning of the chromosomes for regional reference. Samples are grouped together by similarity in pairwise CNA comparison using UPGMA and are labeled with the corresponding MOE. The bottom panel is the summary of chromosomes with the most recurrent changes. Chromosomes with significantly more recurrent changes in CAP are represented by the yellow, and cancer by the olive green color. The stepwise MOEs are shown in red, parallel in yellow, the combination of two in orange, neutral in gray, and eruptive in blue.

Deletions in the p arm of chromosome 1 are significantly recurrent in CFP and CAP compared to other chromosomes with p-values of $8x10^{-4}$ and $2.4x10^{-6}$, respectively. This indicates that it is possibly an adenoma-specific marker that is subclonal so that the deletion is outcompeted by another subclone without the deletion at the cancer stage. Duplications in chromosome 7 are specifically recurrent in CAP ($p=4.2x10^{-5}$) and cancer ($p=4.8x10^{-7}$). Deletions in chromosomes 17 and 18 are also significantly recurrent compared to other chromosomes in CAP with a p-value of $3.3x10^{-5}$ for both and in cancer with p-values of $6.7x10^{-6}$ and $4.8x10^{-7}$ respectively. These chromosomes with repeated alterations compared to other chromosomes in CAP and cancer demonstrate that these copy number alterations may be indicative of malignancy. Similarly, duplications in chromosome 13 ($p=1.4x10^{-6}$), and deletions in chromosome 21 ($p=2.6x10^{-5}$) and 22 ($p=1.6x10^{-5}$) may be an indicator that the sample is benign as they are most recurrently present only in CFP. It is also possible that these are potential markers that signify preventative nature against cancer. For example, the large duplication in chromosome 13 could have a protective effect against the duplication on chromosome 7 as long as the deletion on chromosome 17 is not present. Thus, our findings suggest the possibility for these markers to be used for potential early detection of the malignant polyps.

Utility of exome sequencing in MOE classification

To determine if our criteria for MOE classification are also applicable to exome sequencing data, the 2D plots for case A03 and A09 were re-created based on AF of the somatic SNVs found in coding regions only (**App. Fig. 17**). All features of the 2D plot from the whole genome sequencing data were observable in the new 2D plot. Those cases with fewer number of SNVs may lose some features of the AF distribution simply due to the lower number of SNVs in the exome compared to the genome, but the criteria appears to be robust to exome sequencing overall. Similarly, the differential occurrence of aneuploidies may still be observable despite the less definitive amplification and CNAs from the exome. Our criteria should be applied to an actual exome sequencing data to determine whether the MOE classification can still apply to exome sequencing.

DISCUSSION

There have been several studies modeling the dynamic process of neoplastic transformation based on the spatial characterization of cancer by multi-region sampling as well as temporal progression of cancer by comparing the primary cancer sample to metastases [34,44]. Traditionally intra-tumor heterogeneity from multiple spatially distinct regions of a cancer at one point in time or possibly in corresponding cancer recurrence tissues has been one means to evaluate the clonal history of the tumor and create phylogenetic trees that depict cancer evolution [45,46]. However, these studies do not take into account information of intermediate clones that do not persist through malignant transformation. Even if they do, they are often exclusively based on unrelated polyps that have not developed into cancer or cancers in which the presumed polyp of origin is no longer present, i.e., without directly evaluating the molecular transformation

of normal colon through polyp to cancer in the same person. In addition to the studies on spatial characterization of cancer, single cell sequencing studies have contributed in the in-depth analyses of the clonal lineage in carcinogenesis with the ability to call somatic mutations that are missed due to their low AFs [47,43]. Nevertheless, these approaches still fail to address the fundamental limitation in the study design of explaining cancer evolution from CRC sample alone.

In this study, we used the cancer adjacent polyp as a snapshot of the CRC clonal lineage to infer the pre-cancer time course. Numerous studies have reported that the remnant features of primary tumors in their recurrent metastatic tumors, suggesting that the determination of whether the tumor will metastasize could possibly come early during the primary tumor progression [44,46,48]. Similarly, we posit that the cancer's MOE arises in polyp compartment, which precedes the presence of cancer.

We defined a set of criteria that is based on the information from the presence of both cancer and its matching polyp in order to classify each case into one of four MOEs. Our defined criteria are almost universal as we were able to classify all but two cases into a distinct MOE with no ambiguity. Complications in classifying the two cases could stem from those cases not belonging to any of the established four MOEs. It is possible that the two cases have undergone an extremely rare and not well-understood MOE such as CRC with polyclonal origin [43]. For every mode, we found at least one case corresponding to it. Approximately 70% of CRCs in our dataset evolved in a stepwise or a parallel fashion, although we should note that all cases are expected to exhibit features of parallel evolution, given the nature of our experiment i.e. existence of matched polyp and CRC. We also found that the non-aggressive cases in which patients with stage II and III CRC

survived their cancer exhibited all four stepwise, parallel, neutral, or eruptive MOE. This implies that MOE does not determine or at least is not the only determinant of clinical aggressiveness.

Our data on the relationship between the MOE and the somatic substitution mutations show that the patterns in the somatic substitutions were more significantly influenced by the underlying mutational mechanisms than by their MOEs. It also indicates that several different mutational mechanisms can lead to the same MOE and two different MOEs can have similar mutational mechanisms. Similarly, CNA patterns did not directly correlate with the MOE. This indicates that mutations in specific genes or pathways, and sequential order of deletions and duplications could determine the MOE of a CRC. While it is possible that factors other than the mutational mechanism – such as somatic variants in particular pathways, germline variants, and perhaps the microbiome – that can determine the MOE of a CRC, it is also possible that the relationship between MOE and mutational mechanism simply could not be found given our small sample size. Thus our finding warrants further study in understanding the relationship between the two.

Currently, the MOE assignment requires both the cancer and the matching polyp, so the CFPs were not assigned the MOE. Yet, if the MOE could be determined at a polyp stage solely based on the transformation from normal colon cell to the polyp, we could predict the clinical phenotype of the patient as well as the mutational architecture of any later stages. Our analysis of the mutational spectra illustrates that the CFPs cluster among the CAPs. Similarly, CFPs and CAPs seemed to share some of the recurrent copy number alterations. This comparison between the CFPs and CAPs accounts for the potential bias

in the representation of CAPs and leaves open the possibility of assigning the MOEs at the polyp stage in order to determine whether it progresses to cancer or not.

CONCLUSION

Overall, our study is the first to define specific criteria to link the MOEs to mutational characteristics in patient cases using cancer and its residual polyp of origin. Just as carcinogenesis models are relevant across different cancers, this MOE classification of CRC may apply to other cancer types with premalignant stages including PanINs in pancreatic cancer and ductal hyperplasia that precedes breast cancer [49,50]. Use of MOE in the study of neoplastic transformation promises to provide additional insight into the genomic landscape of cancer as our classification signifies a characterization of cancer that is vastly different from the conventional genomic profiling using somatic SNVs and CNAs. Combining single cell analysis and intra-tumoral/intra-polyp heterogeneity approaches focusing on the more precise tracking of cancer evolution from the early, perhaps polyp-specific event initiating neoplastic transformation will most likely provide additional insights into the details of malignant transformation in CRC. This insight, along with the relevance of CRC MOE modeling in other cancer types, may better our understanding of carcinogenesis in order to improve prognostication and to develop treatments targeted at the most relevant molecular events that drive both malignant transformation and ultimately progression.

BIBLIOGRAPHY

1. Siegel RL, Miller KD, Fedewa SA, Ahnen DJ, Meester RGS, Barzi A, et al. Colorectal Cancer Statistics , 2017. 2017;67:177–93.

2. Hess KR, Varadhachary GR, Taylor SH, Wei W, Raber MN, Lenzi R, et al. Metastatic patterns in adenocancer. Cancer. 2006;106:1624–33.

3. Haggar F a, Boushey RP, Ph D. Colorectal Cancer Epidemiology : Incidence , Mortality , Survival , and Risk Factors. Clin. Colon Rectal Surg. 2009;6:191–7.

4. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61:759–67.

5. Kershaw SK, Byrne HM, Gavaghan DJ, Osborne JM. Colorectal cancer through simulation and experiment. IET Syst. Biol. [Internet]. 2013;7:57–73. Available from: http://digital-library.theiet.org/content/journals/10.1049/iet-syb.2012.0019

6. Boland RC, Goel A. Microsatellite Instability in Colorectal Cancer. Gastroenterology. 2010;138:2073–87.

7. Ribic CM, Sargent DJ, Moore MJ, Thibodeau SN, French AJ, Goldberg RM, et al. Tumor Microsatellite-Instability Status as a Predictor of Benefit from Fluorouracil-Based Adjuvant Chemotherapy for Colon Cancer. New English J. Med. 2003;349:247–57.

8. Goel A, Xicola RM, Nguyen T, Doyle BJ, Vanessa R, Bandipalliam P, et al. NIH Public Access. 2011;138:2044–58.

9. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, et al. Feasibility of screening for Lynch syndrome among patients with colorectal cancer. J. Clin. Oncol. 2008;26:5783–8.

10. Vasen H, Wijnen J, Menko F, Kleibeuker J, Taal B, Griffioen G, et al. Cancer risk in families with hereditary nonpolyposis colorectal cancer diagnosed by mutation analysis. Gastroenterology. 1996;110:1020–7.

11. Scholer-Dahirel A, Schlabach MR, Loo A, Bagdasarian L, Meyer R, Guo R, et al. Maintenance of adenomatous polyposis coli (APC)-mutant colorectal cancer is dependent on Wnt/ -catenin signaling. Proc. Natl. Acad. Sci. [Internet]. 2011;108:17135–40. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1104182108

12. Markowitz S, Bertagnolli M. Molecular Basis of Colorectal Cancer. N. Engl. J. Med. [Internet]. 2009;361:2449–60. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843693/pdf/nihms-177087.pdf

13. Heppner Goss K, Groden J. Biology of the adenomatous polyposis coli tumor suppressor. J. Clin. Oncol. 2000;18:1967–79.

14. Soussi T, Wiman KG. Shaping Genetic Alterations in Human Cancer: The p53 Mutation Paradigm. Cancer Cell. 2007;12:303–12.

15. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature [Internet]. Nature Publishing Group; 2013;502:333–9. Available from: http://www.nature.com/doifinder/10.1038/nature12634

16. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic alternations during colorectal-tumor development. N. Engl. J. Med. 1988;319:525–32.

17. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. Genomics [Internet]. Elsevier Inc.; 2008;92:255–64. Available from: http://dx.doi.org/10.1016/j.ygeno.2008.07.001

18. Shendure J, Ji H. Next-generation DNA sequencing. Nat. Biotechnol. [Internet]. 2008;26:1135–45. Available from: http://www.nature.com/doifinder/10.1038/nbt1486

19. Mitra RD, Church GM. In situ localized amplification and contact replication of many individual DNA molecules. Nucleic Acids Res. [Internet]. 1999;27:34e–34. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/27.24.e34

20. Lee H, Gurtowski J, Yoo S, Nattestad M, Marcus S, Goodwin S, et al. Third-generation sequencing and the future of genomics. bioRxiv [Internet]. 2016; Available from: http://biorxiv.org/content/early/2016/04/13/048603.abstract

21. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008;24:133–41.

22. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. [Internet]. 2009;10:57–63. Available from: http://www.nature.com/doifinder/10.1038/nrg2484

23. Furey TS. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2012;13:840–52. Available from: http://www.nature.com/doifinder/10.1038/nrg3306

24. Futreal P a, Kasprzyk a, Birney E, Mullikin JC, Wooster R, Stratton MR. Cancer and genomics. Nature [Internet]. 2001;409:850–2. Available from: http://www.nature.com/nature/journal/v409/n6822/pdf/409850a0.pdf

25. Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br. J. Cancer [Internet]. 2004;91:355–8. Available from: http://www.nature.com/doifinder/10.1038/sj.bjc.6601894

26. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat. Commun. [Internet]. Nature Publishing Group; 2015;6:8971. Available from: http://www.nature.com/doifinder/10.1038/ncomms9971

27. Andor N, Graham TA, Jansen M, Xia LC, Aktipis CA, Petritsch C, et al. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. Nat. Med. [Internet]. 2016;22:105–13. Available from: http://www.nature.com/doifinder/10.1038/nm.3984

28. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. [Internet]. Nature Publishing Group; 2014;15:121–32. Available from: http://www.nature.com/doifinder/10.1038/nrg3642

29. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute

quantification of somatic DNA alterations in human cancer. Nat. Biotechnol. [Internet]. Nature Publishing Group; 2012;30:413–21. Available from: http://www.nature.com/doifinder/10.1038/nbt.2203

30. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. [Internet]. Nature Publishing Group; 2013;31:213–9. Available from: http://www.nature.com/doifinder/10.1038/nbt.2514

31. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Res. 2011;21:974–84.

32. Muzny DM, Bainbridge MN, Chang K, Dinh HH, Drummond JA, Fowler G, et al. Comprehensive molecular characterization of human colon and rectal cancer. Nature [Internet]. Nature Publishing Group; 2012;487:330–7. Available from: http://www.nature.com/doifinder/10.1038/nature11252

33. Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering Signatures of Mutational Processes Operative in Human Cancer. Cell Rep. [Internet]. The Authors; 2013;3:246–59. Available from: http://dx.doi.org/10.1016/j.celrep.2012.12.008

34. Sottoriva A, Kang H, Ma Z, Graham TA, Salomon MP, Zhao J, et al. A Big Bang model of human colorectal tumor growth. Nat. Genet. [Internet]. Nature Publishing Group; 2015;47:209–16. Available from: http://www.nature.com/doifinder/10.1038/ng.3214

35. Kim T-M, An CH, Rhee J-K, Jung S-H, Lee SH, Baek I-P, et al. Clonal origins and parallel evolution of regionally synchronous colorectal adenoma and cancer. Oncotarget [Internet]. 2015;6:27725–35. Available from: http://oncotarget.com/abstract/4834

36. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nat. Genet. [Internet]. Nature Publishing Group; 2016;48:238–44. Available from: http://www.nature.com/doifinder/10.1038/ng.3489

37. Graham TA, Sottoriva A. Measuring cancer evolution from the genome. J. Pathol. 2017;241:183–91.

38. Johnson BE, Mazor T, Hong C, Barnes M, Aihara K, McLean CY, et al. Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. 2014;189:189–94.

39. Druliner BR, Rashtak S, Ruan X, Bae T, Vasmatzis N, O'Brien D, et al. Colorectal cancer with residual polyp of origin: A model of malignant transformation. Transl. Oncol. [Internet]. The Authors; 2016;9:280–6. Available from: http://dx.doi.org/10.1016/j.tranon.2016.06.002

40. Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, et al. Somaticsniper: Identification of somatic point mutations in whole genome sequencing data. Bioinformatics. 2012;28:311–7.

41. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka:

Accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28:1811–7.

42. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25:2283–5.

43. Yu C, Yu J, Yao X, Wu WK, Lu Y, Tang S, et al. Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. Cell Res. [Internet]. Nature Publishing Group; 2014;24:701–12. Available from: http://www.nature.com/doifinder/10.1038/cr.2014.43

44. Brannon AR, Vakiani E, Sylvester BE, Scott SN, McDermott G, Shah RH, et al. Comparative sequencing analysis reveals high genomic concordance between matched primary and metastatic colorectal cancer lesions. Genome Biol. [Internet]. 2014;15:454. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0454-7

45. Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, et al. Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc. Natl. Acad. Sci. [Internet]. 2008;105:13081–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0801523105

46. Navin NE, Hicks J. Tracing the tumor lineage. Mol. Oncol. 2010;4:267–83.

47. Heitzer E, Auer M, Gasch C, Pichler M, Ulz P, Hoffmann EM, et al. Complex tumor genomes inferred from single circulating tumor cells by array-CGH and next-generation sequencing. Cancer Res. 2013;73:2965–75.

48. Zhao Z-M, Zhao B, Bai Y, Iamarino A, Gaffney SG, Schlessinger J, et al. Early and multiple origins of metastatic lineages within primary tumors. Proc. Natl. Acad. Sci. [Internet]. 2016;113:2140–5. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1525677113

49. Notta F, Chan-Seng-Yue M, Lemire M, Li Y, Wilson GW, Connor AA, et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. Nature [Internet]. Nature Publishing Group; 2016;538:378–82. Available from: http://www.nature.com/doifinder/10.1038/nature19823

50. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumour evolution inferred by single-cell sequencing. Nature [Internet]. Nature Publishing Group; 2011;472:90–4. Available from: http://www.nature.com/doifinder/10.1038/nature09807

APPENDIX TABLE OF CONTENTS

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

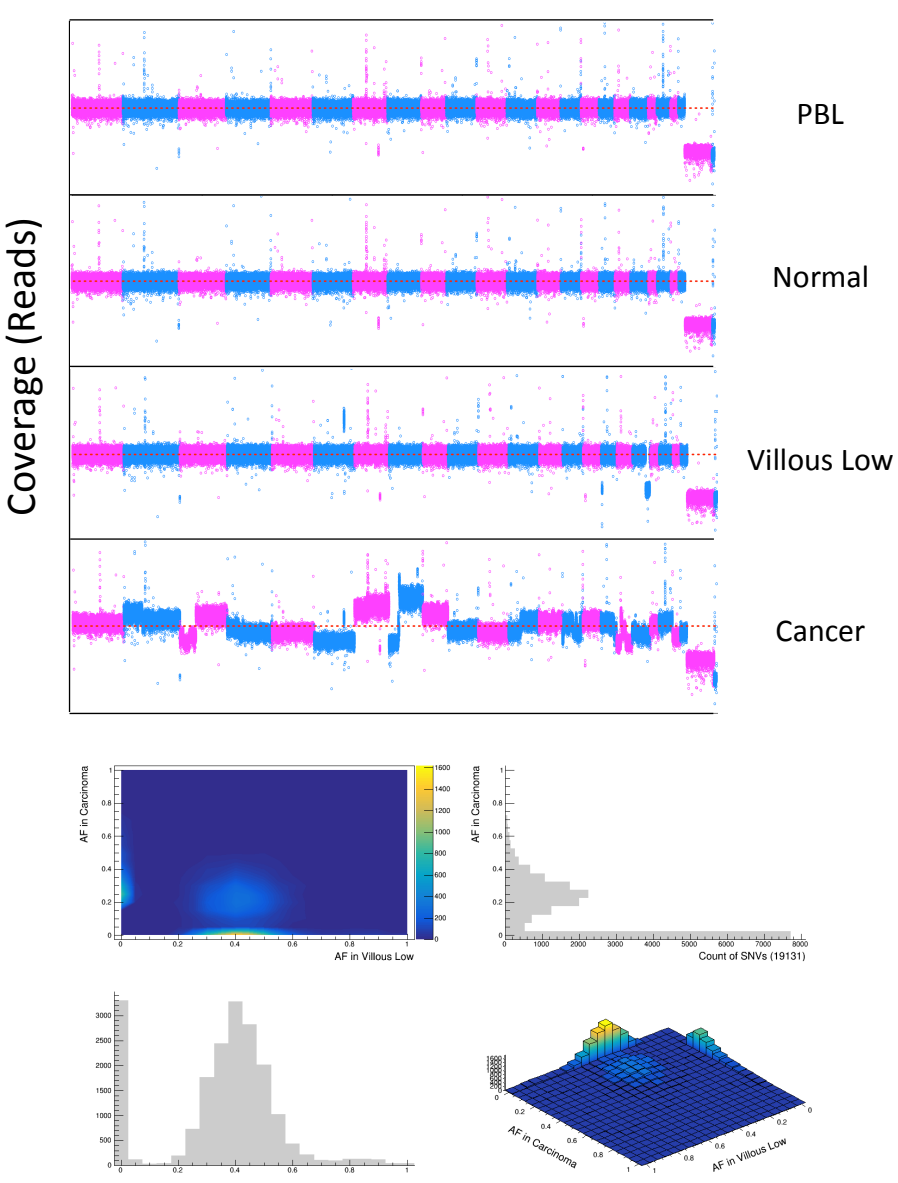
APPENDIX A: Clinical, pathological, and genomic information

Appendix Table 1: **Clinical, pathological, and genomic information for all CRC RPO+ cases**

Microsatellite instable label with high or low level vs. chromosomal instability; modes of evolution; level of cancer purity; aneuploidy level; *APC* mutation percentage in cancer; *TP53* mutation percentage in cancer; *KRAS* mutation percentage in cancer; clinical aggressiveness; and tumor staging.

Cases	MSI/ CIN	MOE	Purity (%)	Aneuploidy Level (%)	APC	TP53	KRAS	Aggressiveness	Stage
A02	CIN	s,p	90	34.44	0	56.7	0	Aggressive	4
A03	CIN	s	80	25.81	37.9	21.9	46.7	Aggressive	4
A07	CIN	s,p	70	27.61	51.9	40.9	31	Aggressive	2
A08	CIN	p	50	31.28	20	44.8	25	Non-aggressive	3
A09	CIN	e	80	27.19	32	69.6	0	Non-aggressive	1
A10	MSI-H	n	70	3.75	0	0	0	Non-aggressive	1
A11	CIN	s	40	22.08	0	16.7; 15.2	0	Non-aggressive	2
A12	MSI-H	s,p	70	2.63	23.3	25	0	Non-aggressive	2
A13	CIN	p	60	8.76	0	2.8	0	Non-aggressive	2
A14	CIN	n	90	3.75	35.7	0	32.1	Non-aggressive	3
A15	CIN	p	50	9.15	20.7	0	26.7	Non-aggressive	3
A16	MSI-H	s	60	4.41	0	0	0	Non-aggressive	2

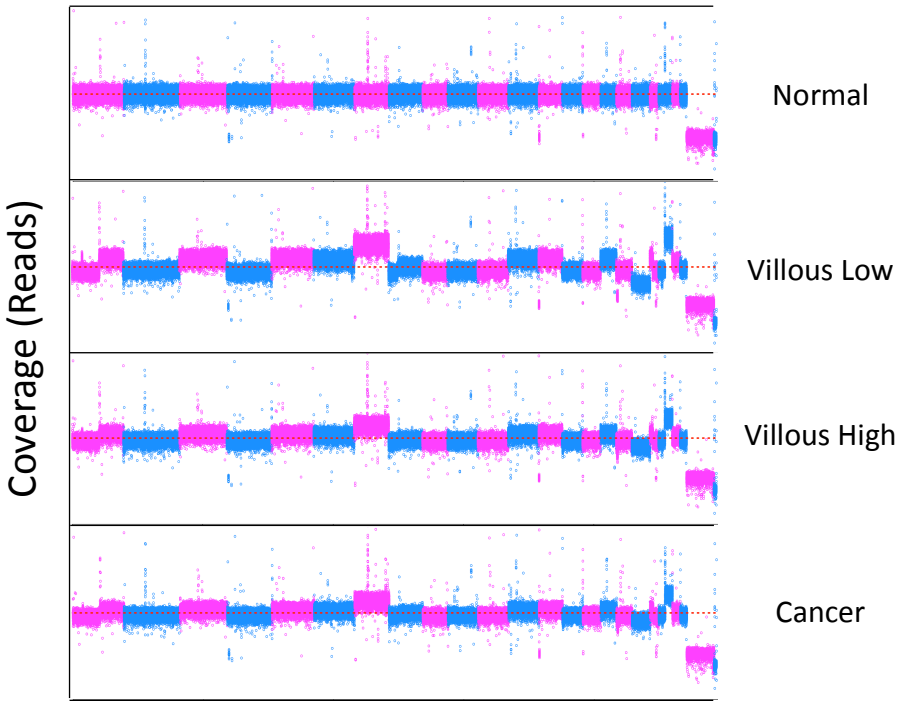
APPENDIX B: Analysis summary for the rest of the cases



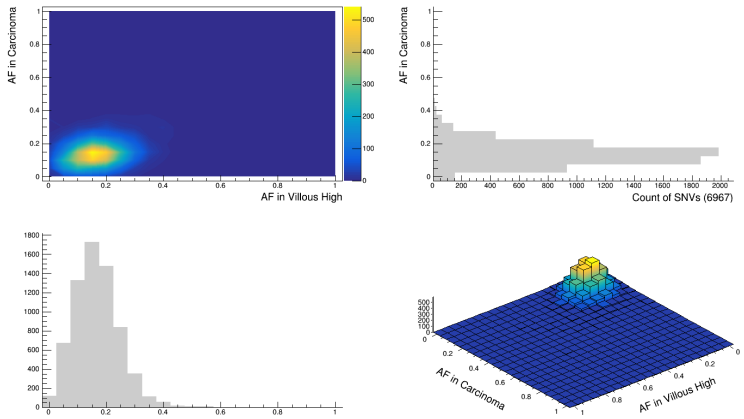
Appendix Figure 1: Analysis summary for case A02, which is either a stepwise or parallel lineage

A) Copy number profile across all chromosomes. Large aneuploidies are observed only at the cancer stage. B) 2D allele frequency distribution is in agreement with the parallel scenario, but stepwise scenario is also possible (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of both the shared SNVs and private SNVs.

A

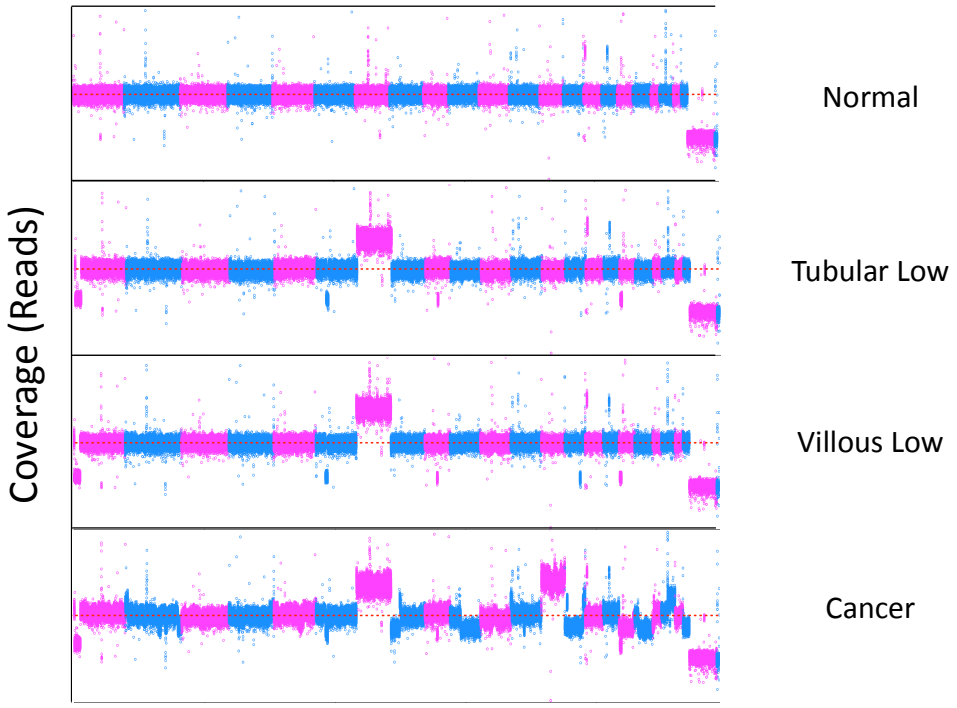


B

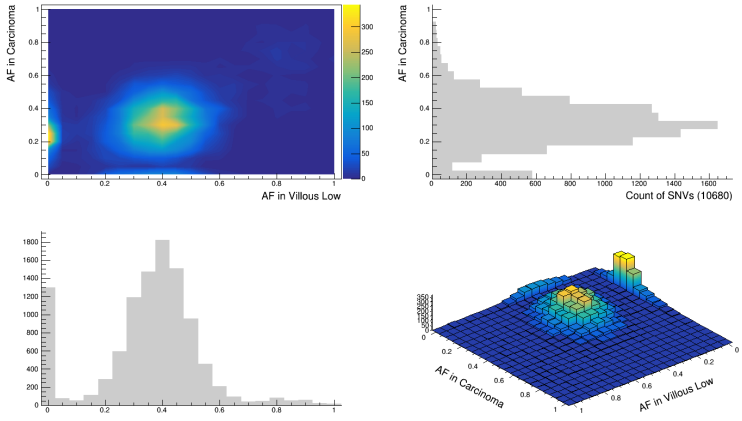


Appendix Figure 2: Analysis summary for case A04, which is likely a neutral lineage A) Copy number profile across all chromosomes. Large aneuploidies are observed before the cancer stage. B) 2D AF distribution is in agreement with the neutral lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates that majority is shared SNVs.

A



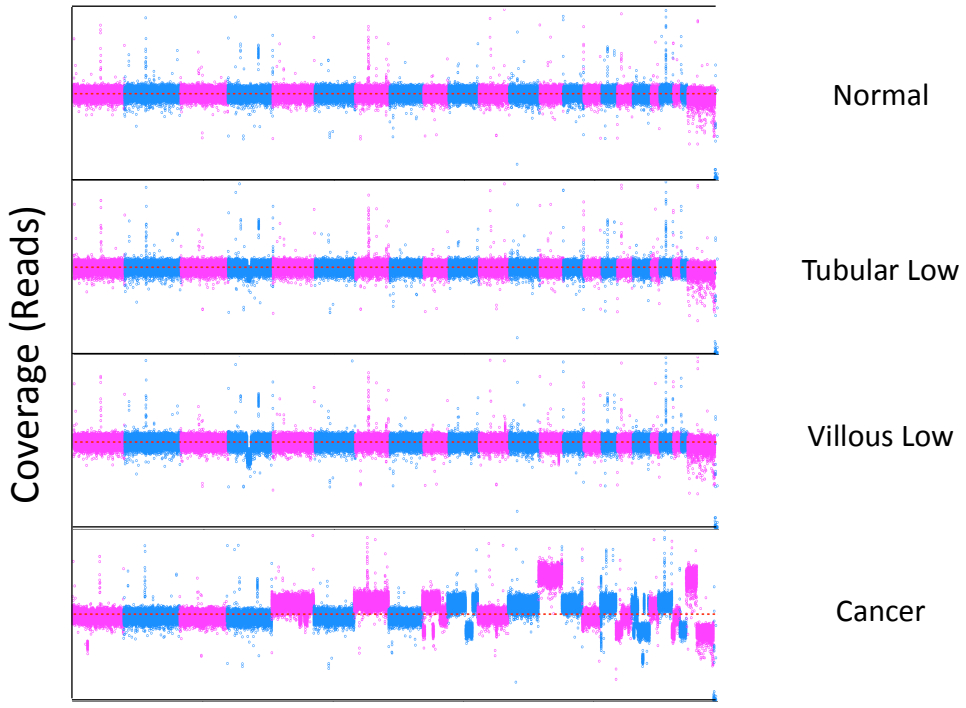
B



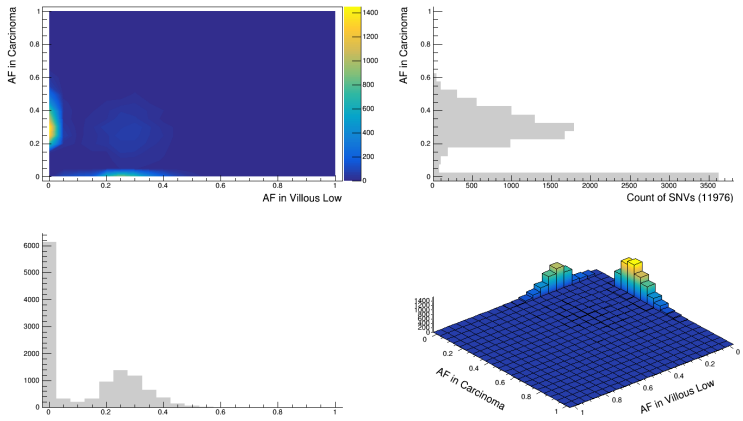
Appendix Figure 3: Analysis summary for case A07, which is either a stepwise or parallel lineage

A) Copy number profile across all chromosomes. The entire chromosome 7 is duplicated early in the progression, but large aneuploidies are not observed until later in the cancer stage. B) 2D allele frequency distribution is in agreement with the parallel scenario, but stepwise scenario is also possible (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of both the shared SNVs and private SNVs.

A

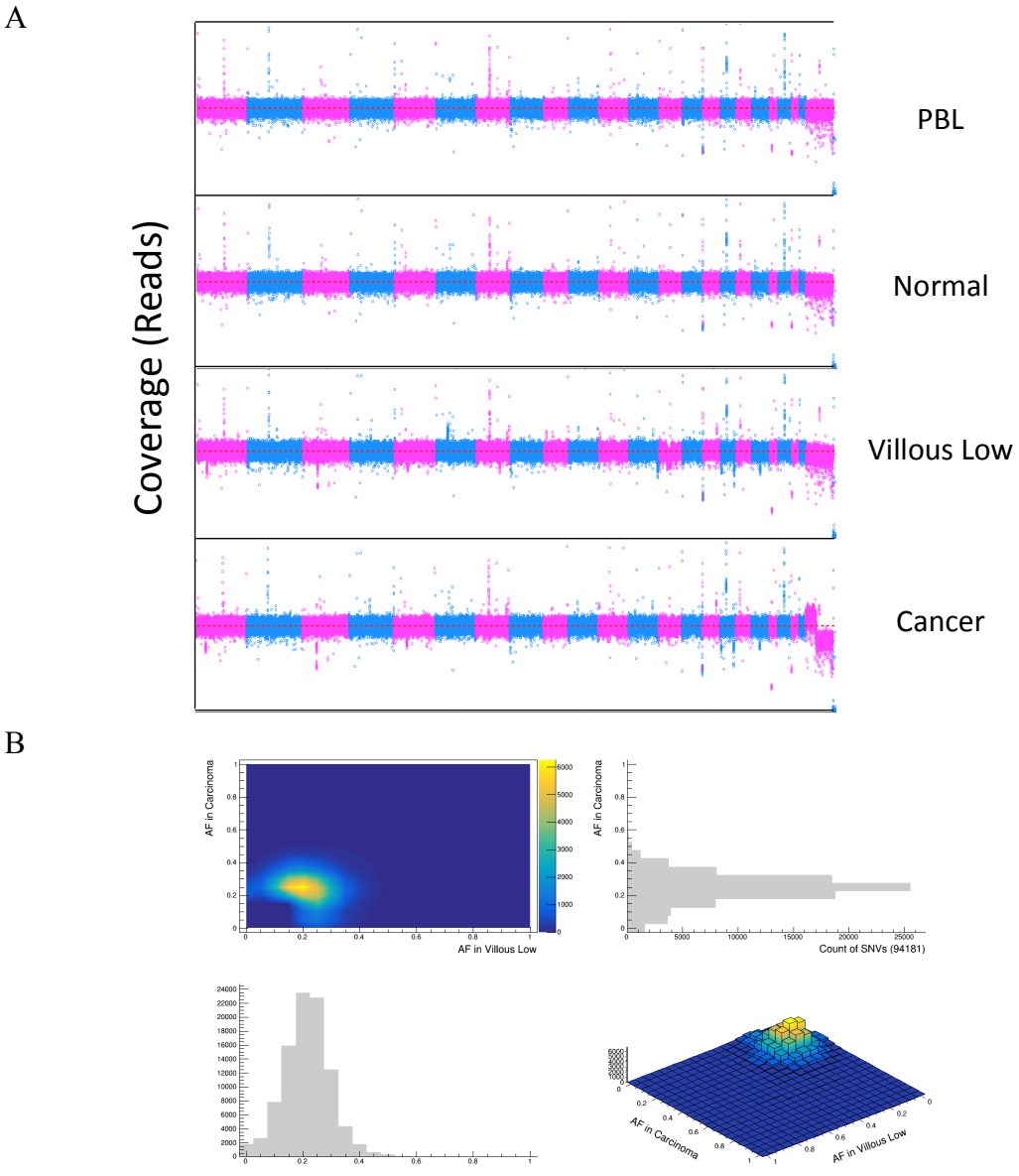


B



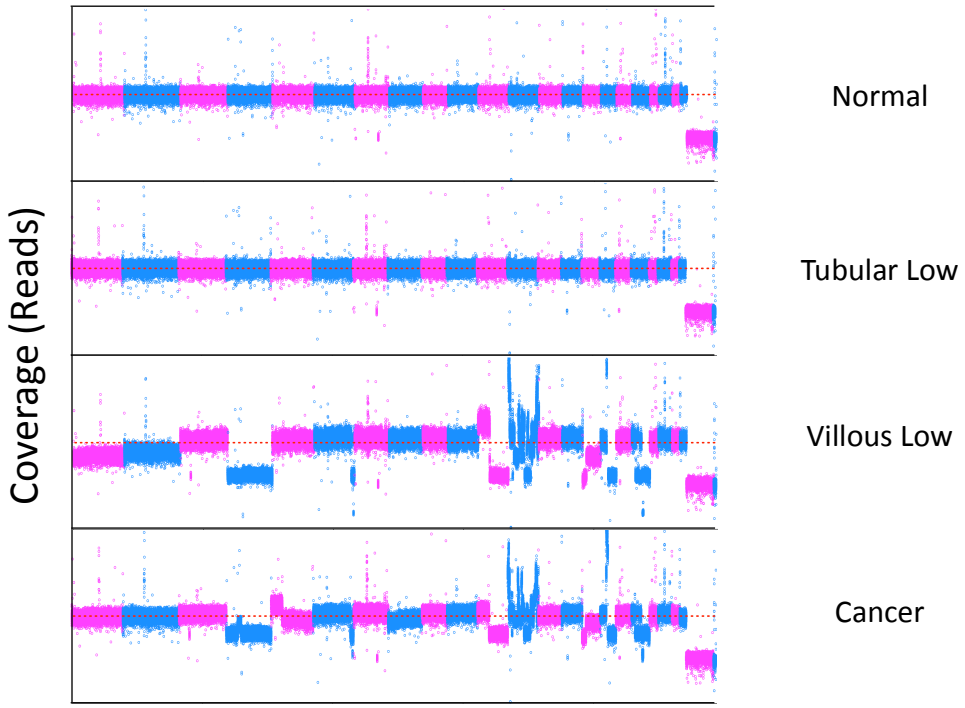
Appendix Figure 4: Analysis summary for case A08, which is likely a parallel lineage

A) Copy number profile across all chromosomes. The deletion on chromosome 4 progressively becomes larger until the cancer stage, at which it is no longer observed; it appears to be subclonal. Large aneuploidies are observed only at the cancer stage. B) 2D AF distribution is in agreement with the parallel lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates that majority is private SNVs.

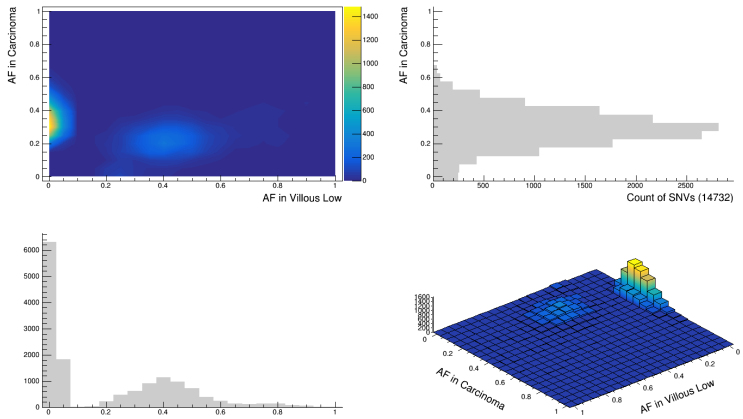


Appendix Figure 5: Analysis summary for case A10, which is likely a neutral lineage A) Copy number profile across all chromosomes. No large aneuploidies are observed. B) 2D allele frequency distribution is in agreement with the neutral lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates that majority is shared SNVs.

A



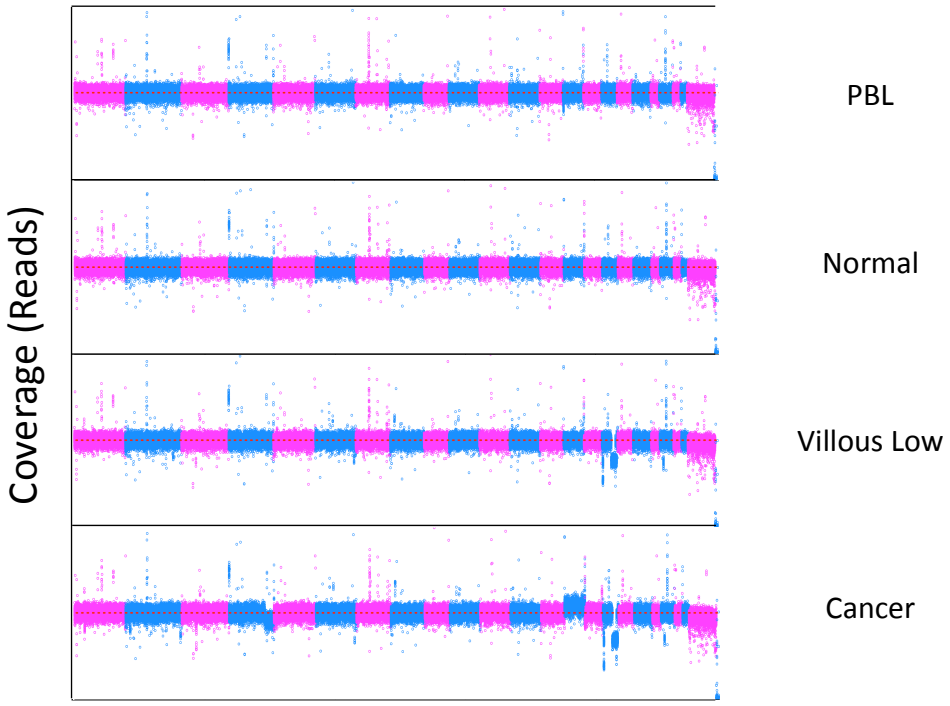
B



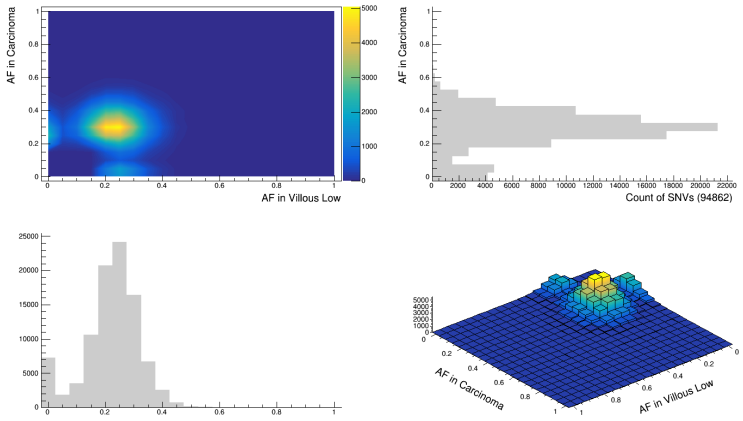
Appendix Figure 6: Analysis summary for case A11, which is likely a stepwise lineage

A) Copy number profile across all chromosomes. Large aneuploidies are observed before the cancer stage. B) 2D allele frequency distribution is in agreement with the stepwise lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of both the shared SNVs and private SNVs.

A



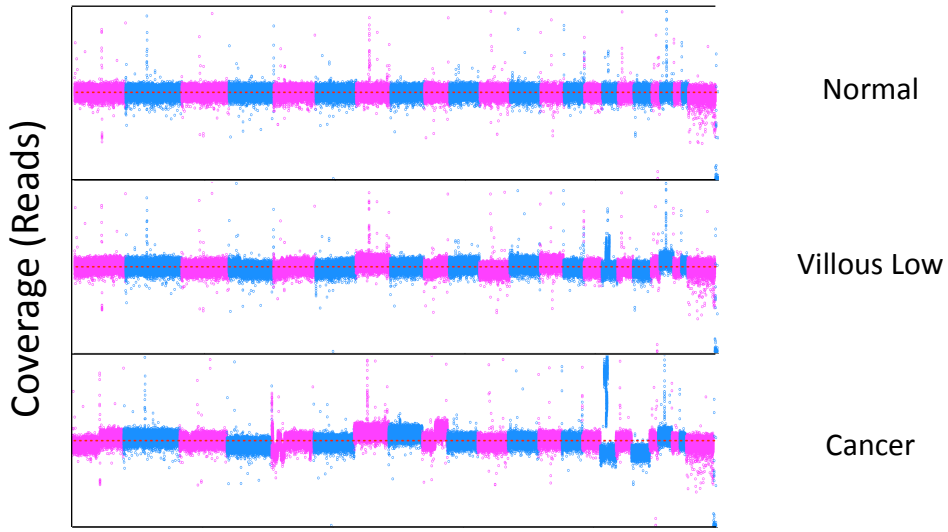
B



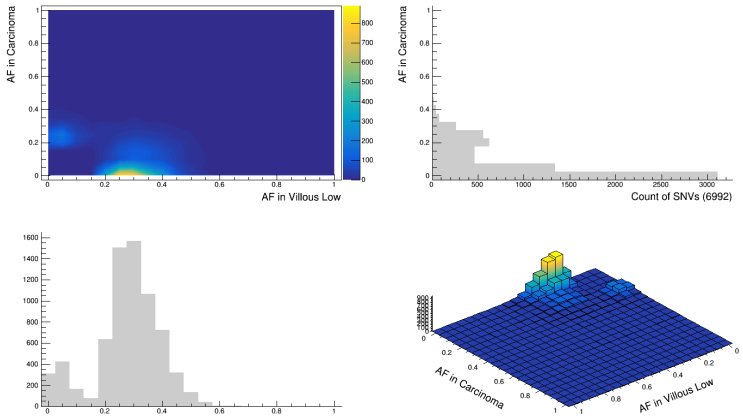
Appendix Figure 7: Analysis summary for case A12, which is either a stepwise or parallel lineage

A) Copy number profile across all chromosomes. No large aneuploidies are observed. B) 2D allele frequency distribution is in agreement with the stepwise scenario, but stepwise scenario is also possible (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of both the shared SNVs and private SNVs.

A



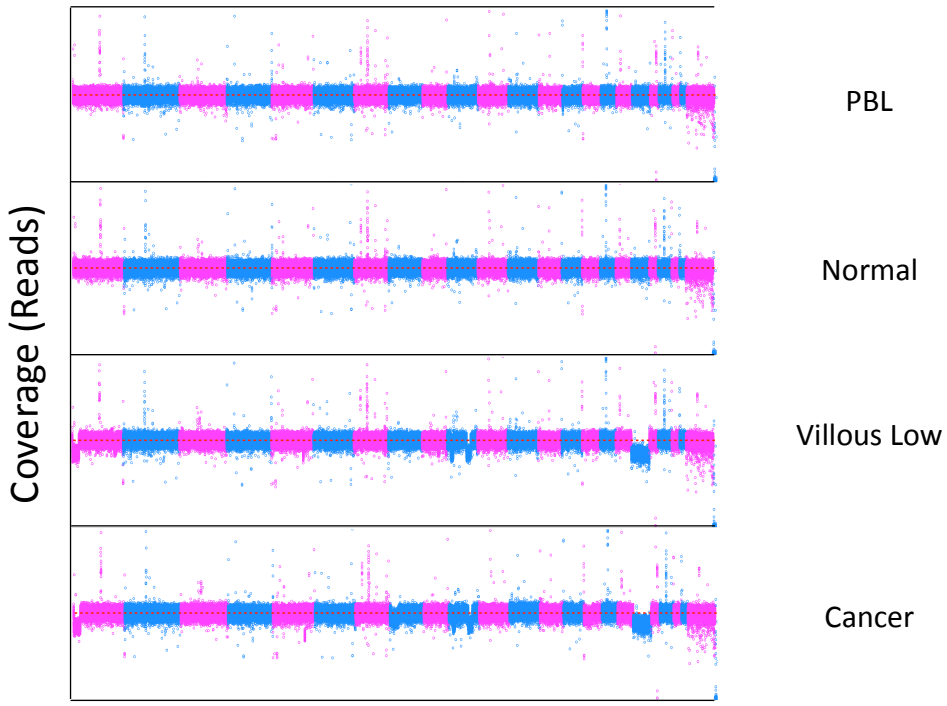
B



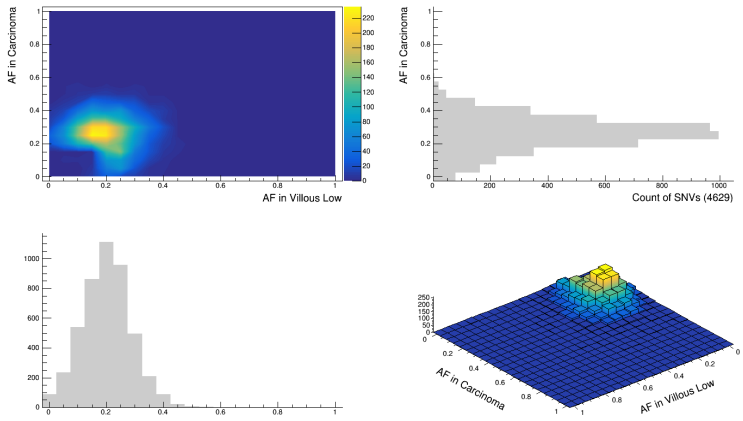
Appendix Figure 8: Analysis summary for case A13, which is likely a parallel lineage

A) Copy number profile across all chromosomes. Large aneuploidies are observed only at the cancer stage. B) 2D allele frequency distribution does not follow any of the scenarios, but based on the number of mutations, it is likely to be a parallel lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates that majority is private SNVs.

A

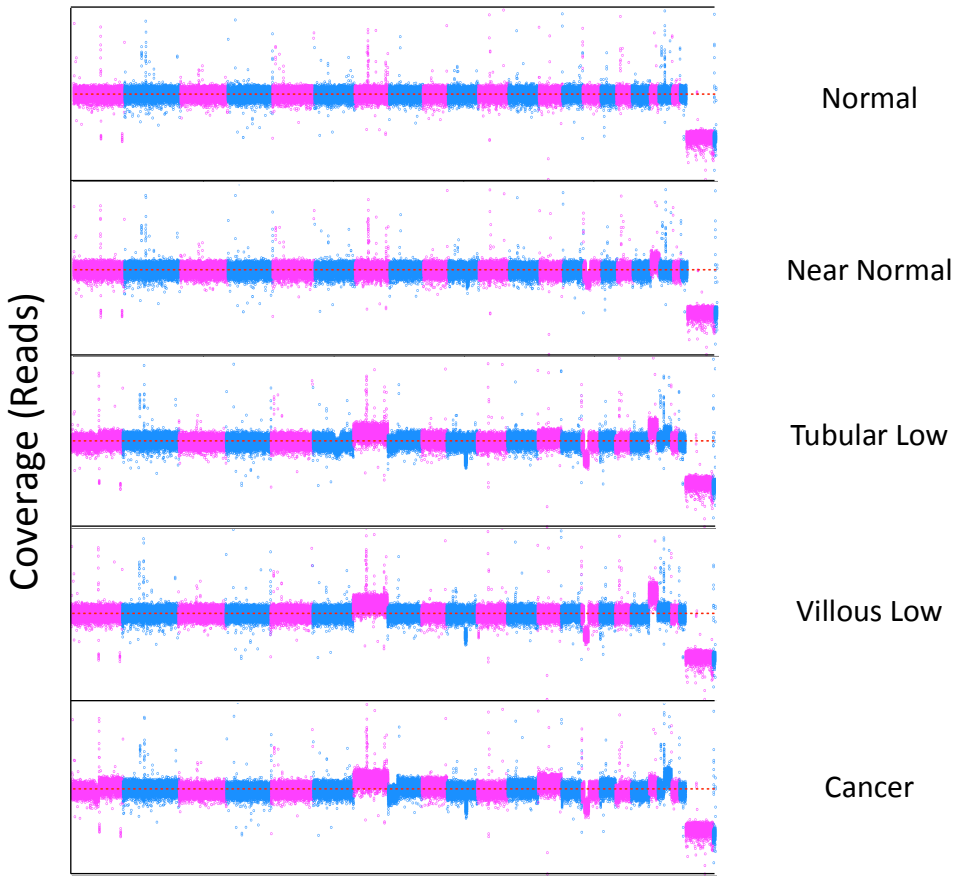


B

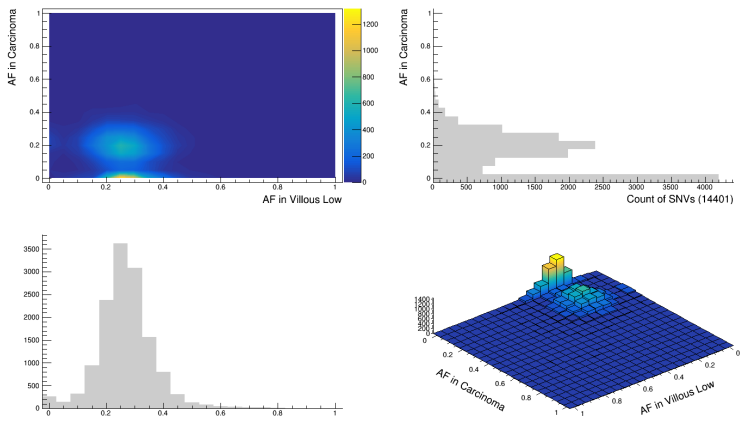


Appendix Figure 9: Analysis summary for case A14, which is likely a neutral lineage A) Copy number profile across all chromosomes. No large aneuploidies are observed. B) 2D AF distribution is in agreement with the neutral scenario (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates that majority is shared SNVs.

A



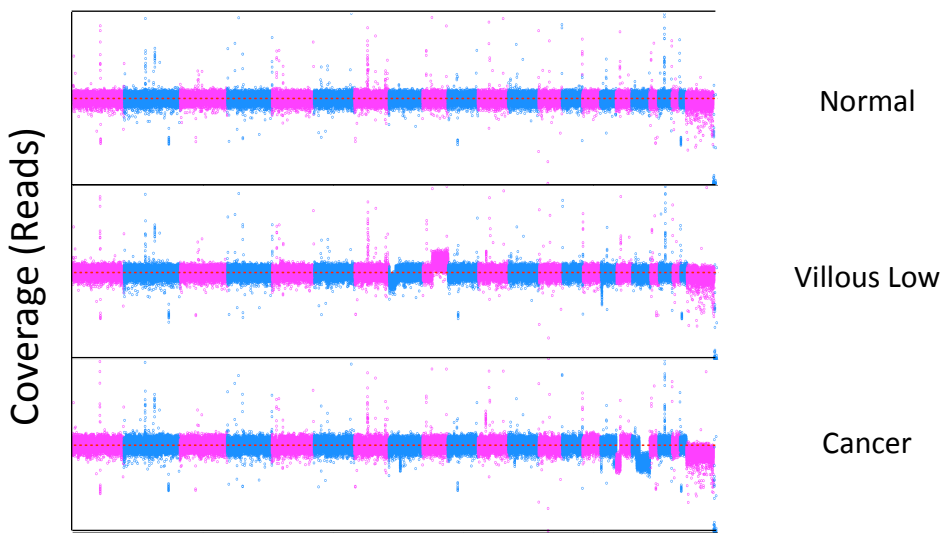
B



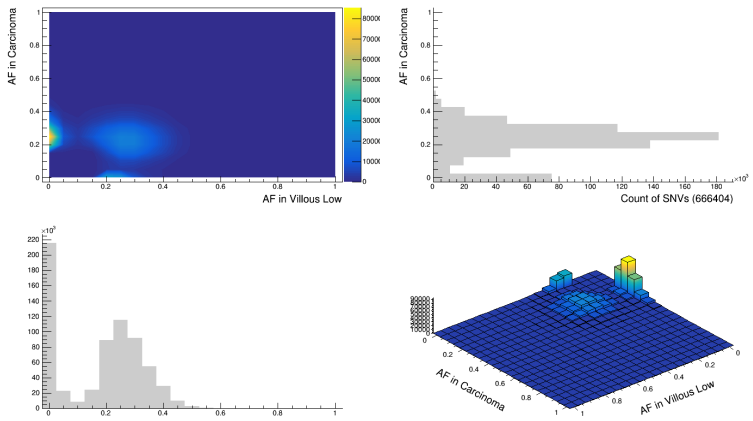
Appendix Figure 10: Analysis summary for case A15, which is likely a parallel lineage

A) Copy number profile across all chromosomes. No large aneuploidies are observed. B) 2D AF distribution is in agreement with the parallel lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of polyp-specific SNVs and shared SNVs.

A



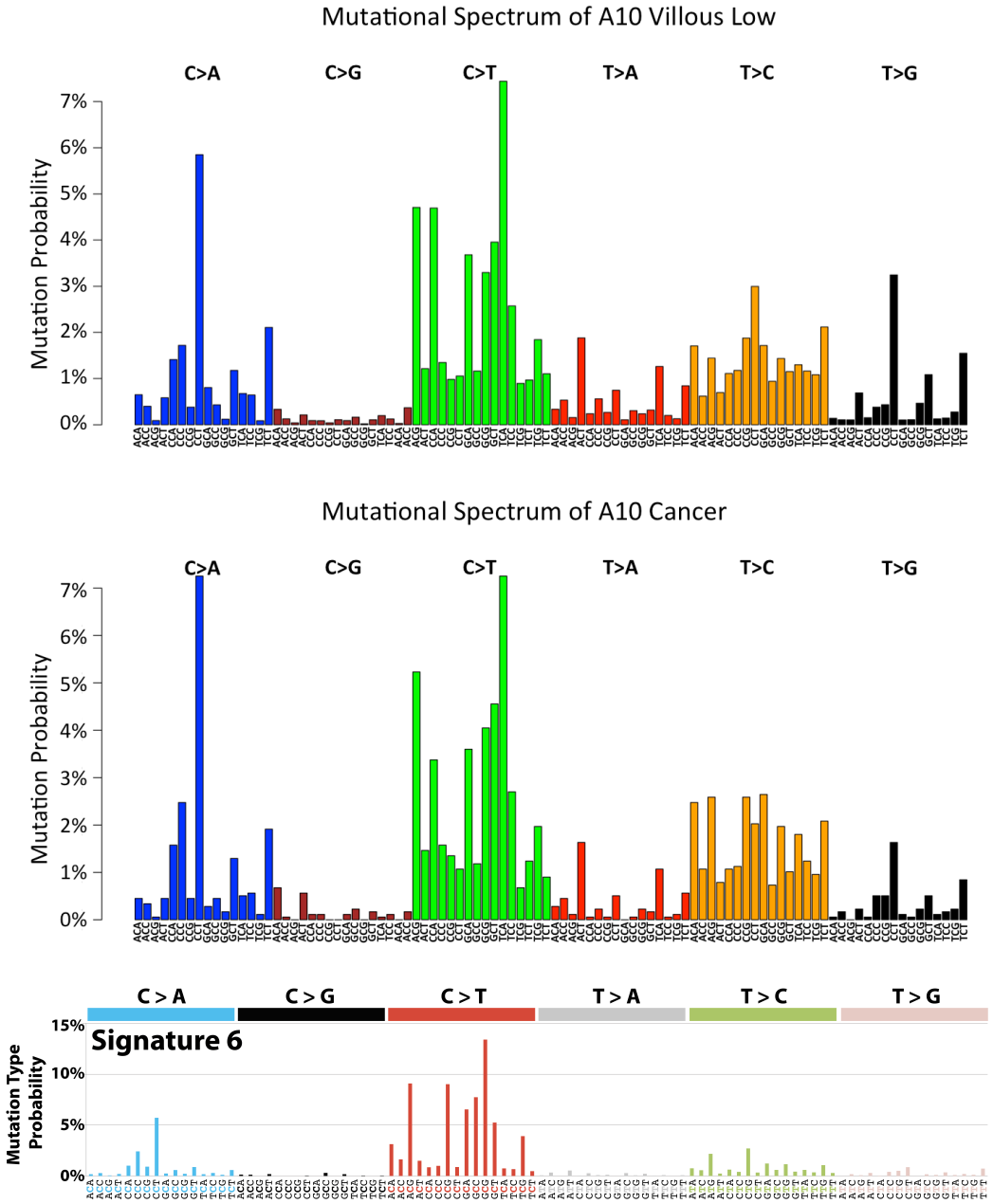
B



Appendix Figure 11: Analysis summary for case A16, which is likely a stepwise lineage

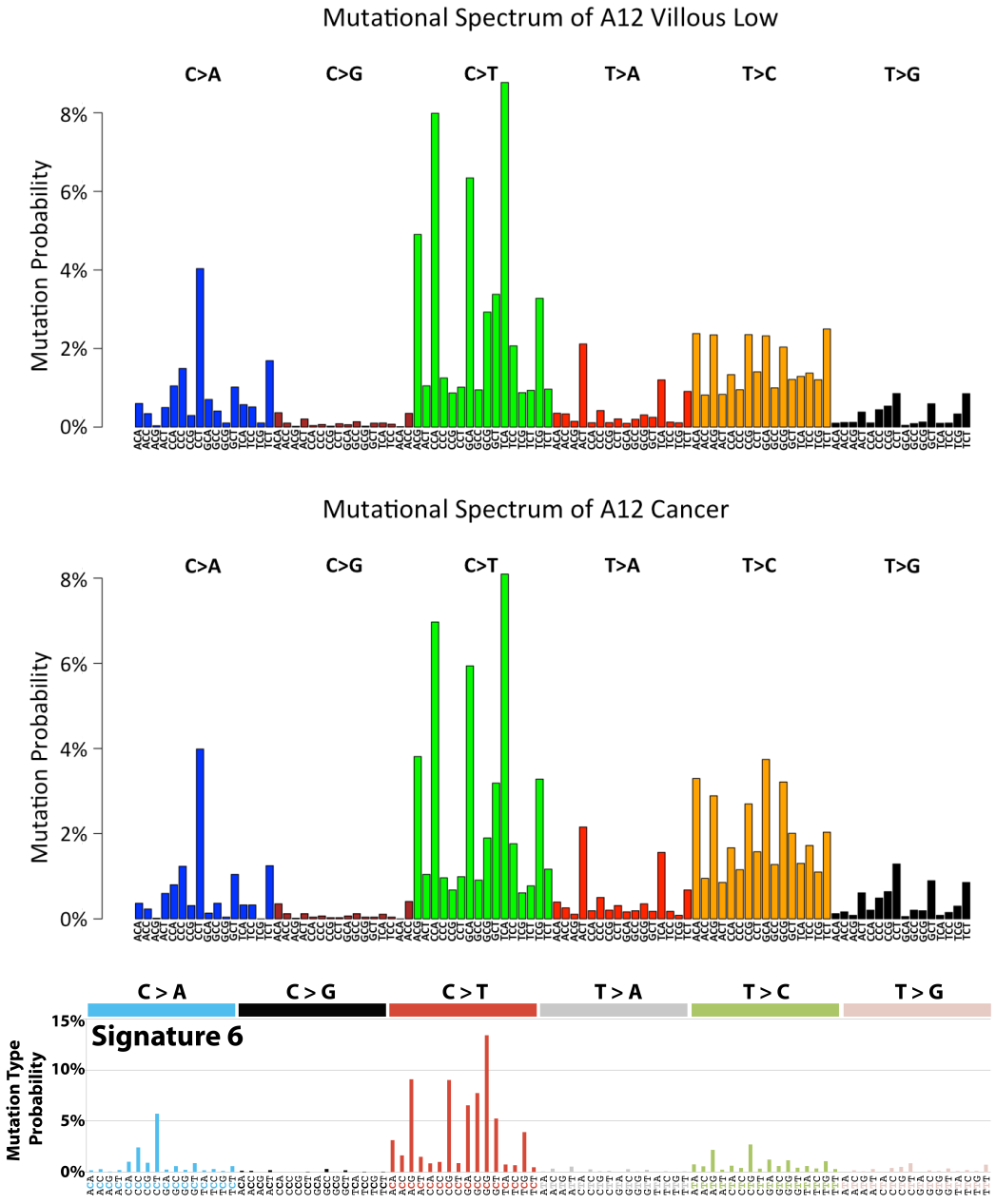
A) Copy number profile across all chromosomes. No large aneuploidies are observed. B) 2D AF distribution is in agreement with the stepwise lineage (see **Fig. 2**). 3D representation of the AF distributions of SNVs is shown on the bottom right. The height of the distributions, which shows the number of mutations, indicates large fraction of both the shared SNVs and private SNVs.

APPENDIX C: Mutational signature analysis of A10, A12, and A16



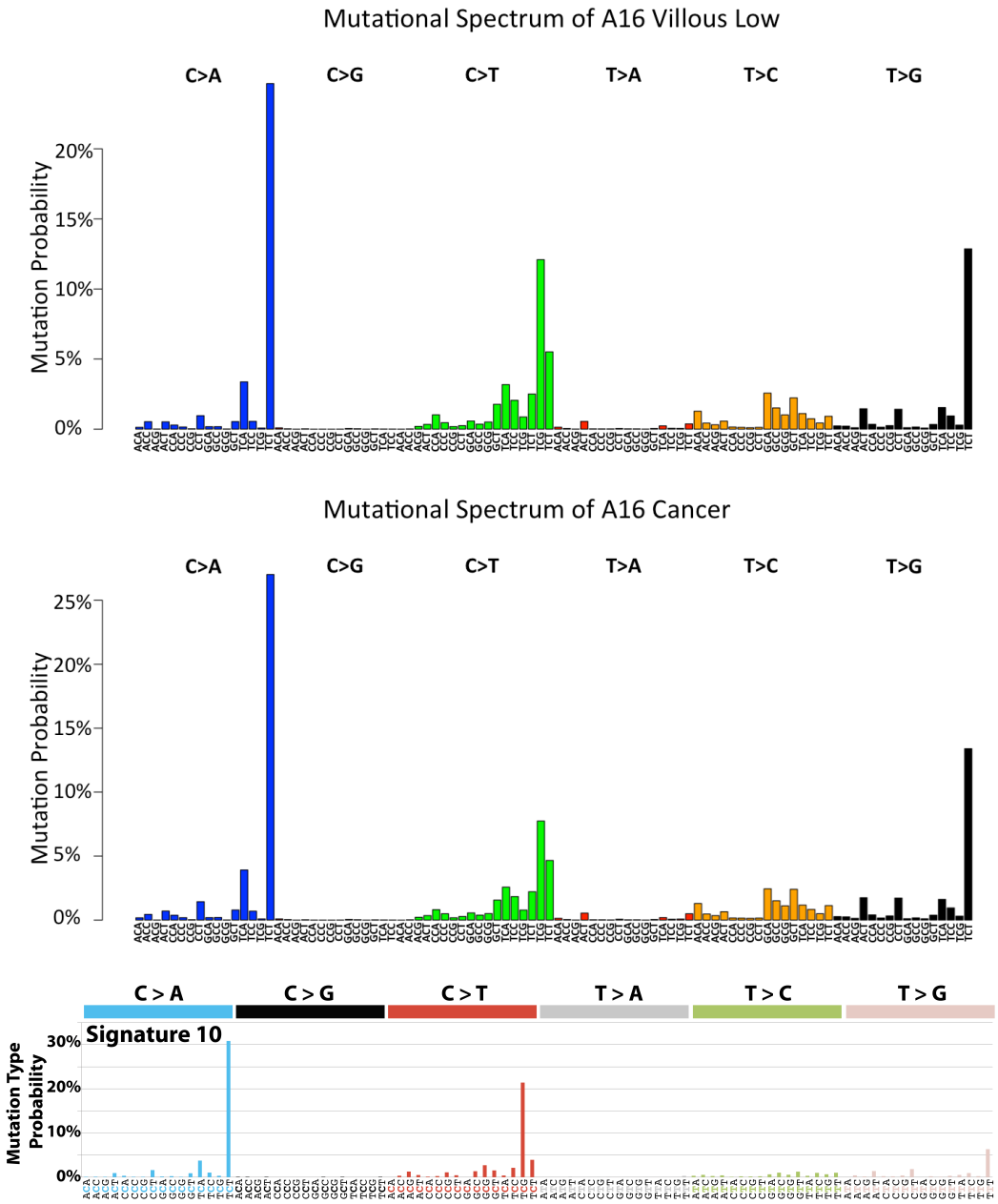
Appendix Figure 12: Mutational spectra of the villous low and cancer in the case A10

The mutational spectra of the case A10 are compared to the mutation signature 6 from the COSMIC database [33]. The significant contributions from C>T and C>A substitutions clearly the resemble the case A10. Signature 6 is presumed to be associated with defective DNA mismatch repair found in MSI tumors (see http://cancer.sanger.ac.uk/cosmic/signatures).



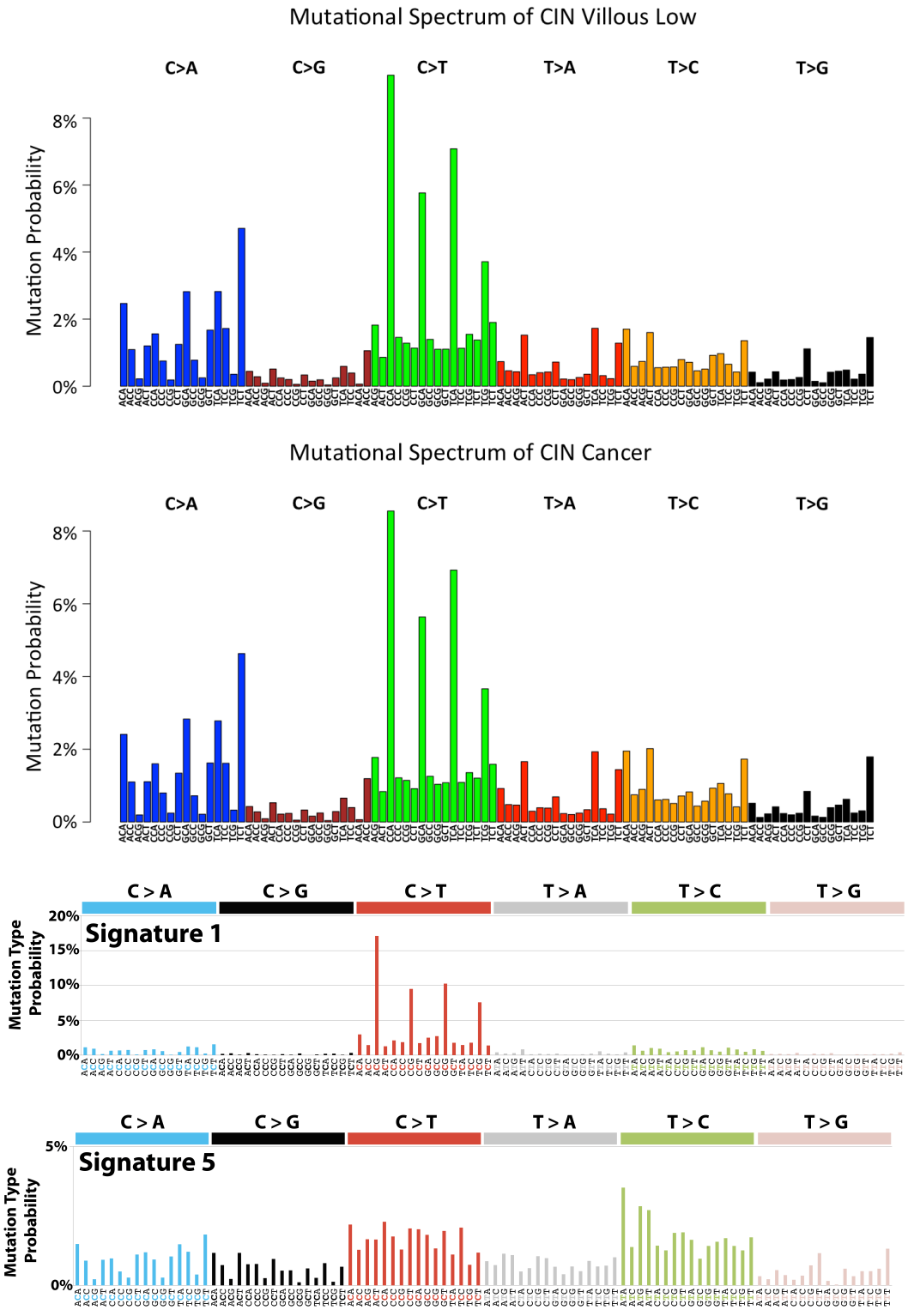
Appendix Figure 13: Mutational spectra of the villous low and cancer in the case A12

The mutational spectra of the case A12 are compared to the mutation signature 6 from the COSMIC database [33]. The significant contributions from C>T and C>A substitutions clearly the resemble the case A12. Signature 6 is presumed to be associated with defective DNA mismatch repair found in MSI tumors (see http://cancer.sanger.ac.uk/cosmic/signatures).



Appendix Figure 14: Mutational spectra of the villous low and cancer in the case A16

The mutational spectra of the case A16 are compared to the mutation signature 10 from the COSMIC database [33]. The significant contributions from C>T and C>A substitutions clearly the resemble the case A16. Signature 10 is presumed to be associated with mutations in POLE gene (see http://cancer.sanger.ac.uk/cosmic/signatures).

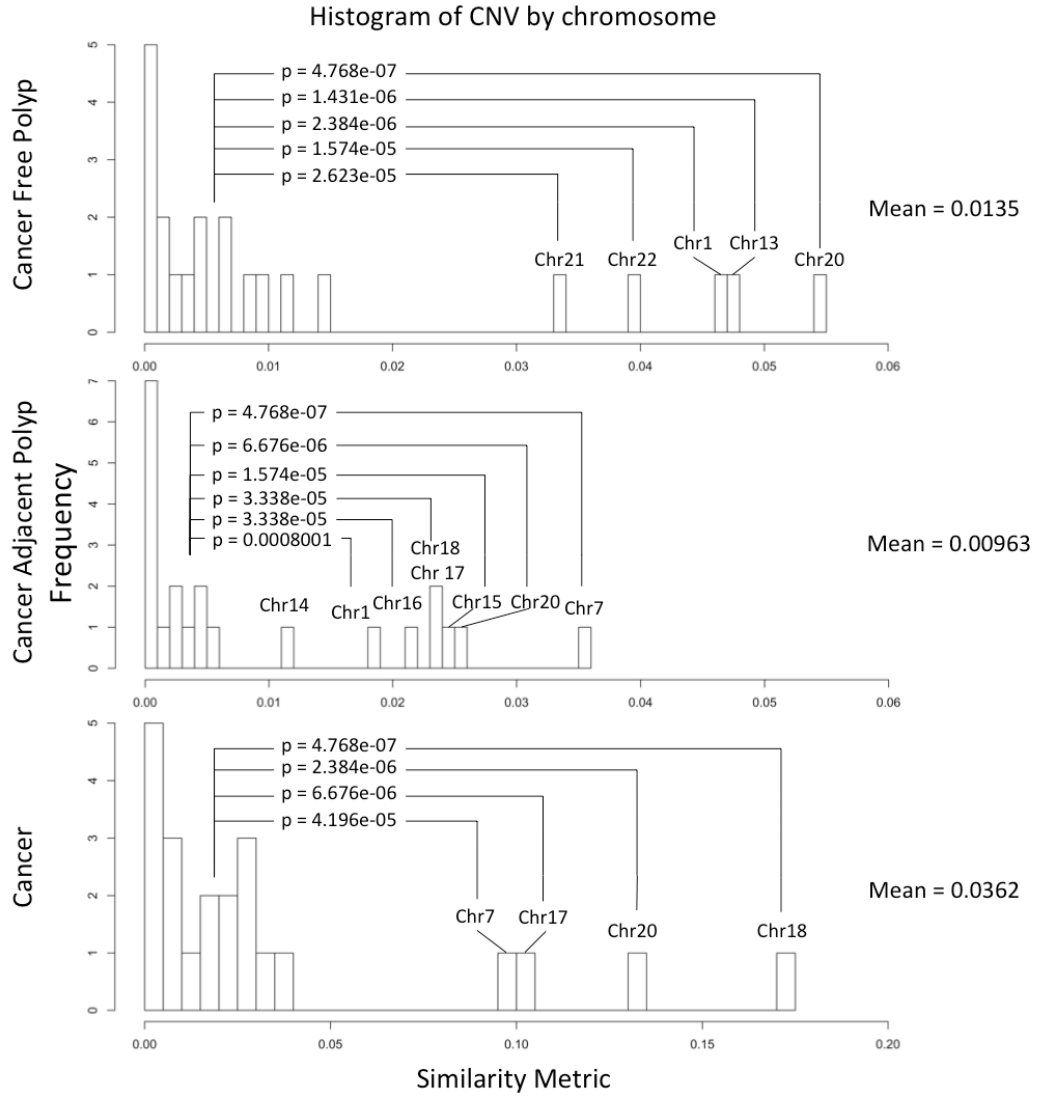


Appendix Figure 15: Mutational spectra of the villous low and cancer in the CIN cases

The mutational spectra of the CIN cases are compared to the mutation signature 1 and 5 from the COSMIC database [33]. The significant contributions from C>A and C>T

substitutions clearly the resemble the CIN cases. Signature 1 is presumed to be associated with sporadic deamination of 5-methylcytosine but signature 5 is not yet known to be associated with particular mechanisms (see http://cancer.sanger.ac.uk/cosmic/signatures).

APPENDIX D: Significantly recurrent CNA by chromosomes for each tissue types

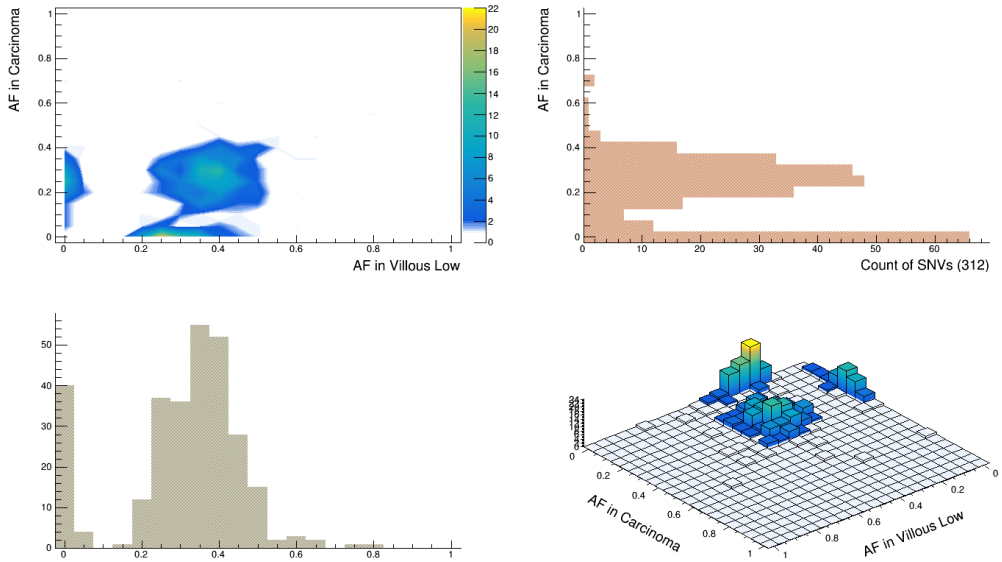


Appendix Figure 16: Histogram of the similarity in the CNA per chromosomes for cancer, CAP, and CFP

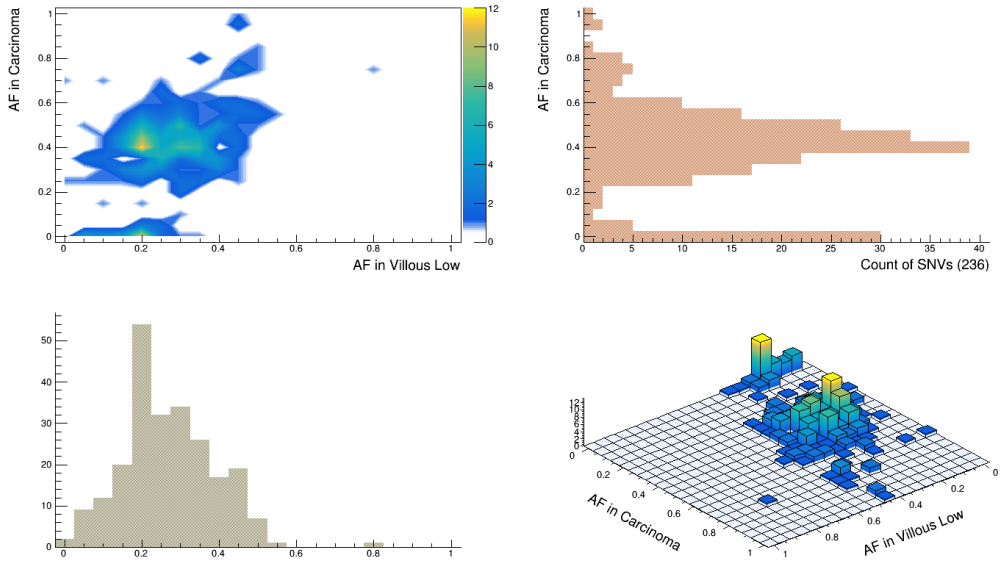
For cancer samples, copy number changes in the chromosomes 18, 20,17, and 7 were significantly more recurrent compared to the other chromosomes. For CAPs, copy number changes in the chromosomes 7, 20,15, 18, 17, 16, and 1 were significantly recurrent compared to the other chromosomes. For CFPs, copy number changes in the chromosomes 20, 13, 1, 22, and 21 were significantly recurrent compared to the other chromosomes.

APPENDIX E: Utility of exome sequencing in MOE classification

A



B



Appendix Figure 17: Allele frequency distributions of SNVs in the coding regions only

A) 2D and 3D AF distributions of SNVs in coding regions for case A03. General shape and the two gaps representing the stepwise evolution remain observable. B) 2D and 3D AF distributions of SNVs in coding regions for case A09. General shape and the single gap in cancer representing the eruptive evolution remain observable.